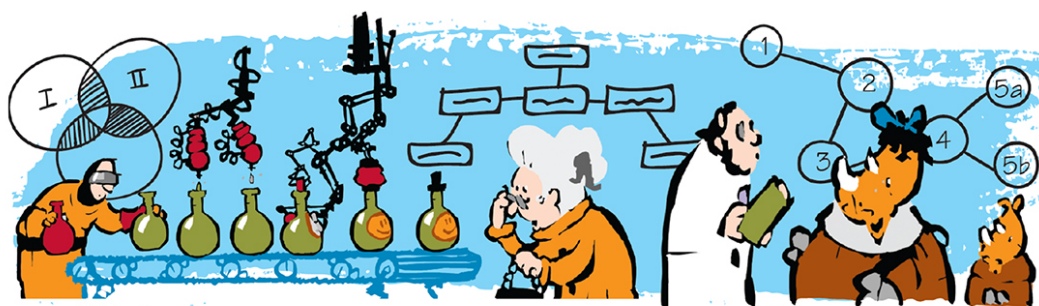


Грейди Клейн и Алан Дебни

СТАТИСТИКА



БАЗОВЫЙ КУРС



В КОМИКСАХ



СТАТИСТИКА

Базовый курс в комиксах

ГРЕЙДИ КЛЕЙН И АЛАН ДЭБНИ



Перевод с английского Ольги Терентьевой

**Москва
«Мани, Иванов и Фербер»
2017**

УДК 311.1
ББК 65.051
К48

Научный редактор Ирина Николаева

*Издано с разрешения
Synopsis Literary Agency с/о THE SYNOPSIS NOA LLP*

На русском языке публикуется впервые

Клейн, Грейди

К48 Статистика. Базовый курс в комиксах / Грейди Клейн, Алан Дебни ; пер. с англ. О. Терентьевой ; [науч. ред. И. Николаева]. — М. : Манн, Иванов и Фербер, 2017. — 240 с.

ISBN 978-5-00100-260-4

Не только полезный, но и веселый курс базовой статистики. Автор и иллюстратор объясняют сложные понятия на простых и забавных примерах, доказывая, что статистика — вокруг нас.

Прочитав эту книгу, вы научитесь собирать данные, делать выборки и проверять гипотезы по любой проблеме — будь то решение о покупке новой машины или подсчет степени взаимной ненависти жителей враждующих планет. Теперь вас не введут в заблуждение показатели средних зарплат по галактике и предвыборные рейтинги, составленные на основе смещенного распределения. Вы узнаете, почему идеальная форма в статистике не менее важна, чем содержание. И в конце концов, получите ответ на важный вопрос, кого огры кидают дальше — эльфов или гномов.

Если же вы захотите мыслить и говорить как статистик, в конце книги вас ждет «Математическая пещера», богатая на формулы и детали.

Книга будет полезна всем, кто хочет познакомиться со статистикой и научиться анализировать данные.

УДК 311.1
ББК 65.051

Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Правовую поддержку издательства обеспечивает юридическая фирма «Вегас-Лекс».

VEGAS LEX

ISBN 978-5-00100-260-4

© THE CARTOON INTRODUCTION TO STATISTICS
by Alan Dabney, illustrated by Grady Klein
Text Copyright © 2013 by Grady Klein and Alan Dabney
Artwork Copyright © 2013 by Grady Klein
Published by arrangement with Hill and Wang,
a division of Farrar, Straus and Giroux, LLC, New York
© Перевод на русский язык, издание на русском языке
ООО «Манн, Иванов и Фербер», 2017

Посвящается Анне, Лиаму и Бенджамину.
Г. К.

Посвящается Эллиотту, Луизе и Нику.
А. Д.

СОДЕРЖАНИЕ

Вступление. Она повсюду ...1

Часть 1. Сбор статистических данных ...15

- 1. Числа ...17**
- 2. Случайные сырые данные ...25**
- 3. Ранжирование ...39**
- 4. Детективная работа ...51**
- 5. Страшные ошибки ...67**
- 6. От выборки к генеральной совокупности ...81**



Часть 2. Поиск параметров ...89

- 7. Центральная предельная теорема ...91**
- 8. Вероятности ...105**
- 9. Статистический вывод ...121**
- 10. Достоверность ...131**
- 11. Они нас ненавидят ...143**
- 12. Проверка гипотез ...161**
- 13. Противостояние ...175**
- 14. Летящие свиньи, плюющиеся пришельцы и петарды ...191**



Заключение. Мыслить как статистик ...205

Приложение. Математическая пещера ...213

Большинство из нас так или иначе имеют дело со статистикой *каждый* день...



Потрясающе!

Одна миска шоколадных шариков содержит **1200%** моей суточной нормы потребления сахара.

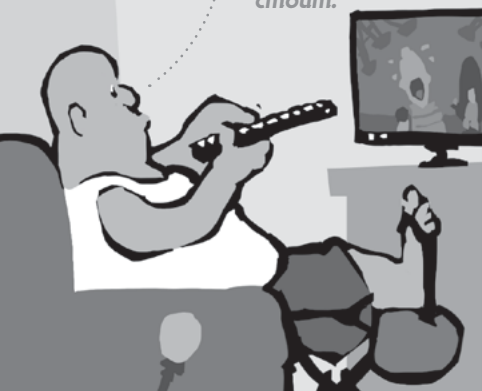


...даже если мы не жонглируем цифрами, зарабатывая себе на жизнь.

Статистику «излучают» наши телевизоры...

Это шоу смотрят **4,8 млн человек!**

Должно быть, оно того стоит.



...она просачивается из телефонов...

В этом месяце вы отправили больше сообщений, чем все население республики Чад.

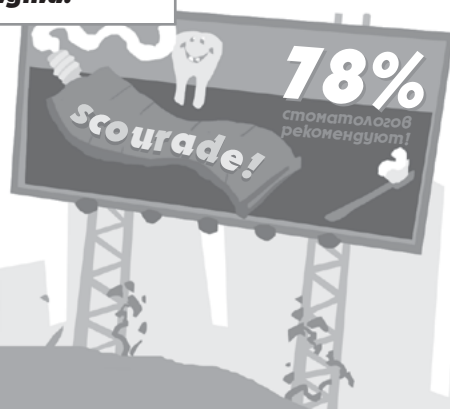


...лбется из радиоприемников...

Согласно опросам, сенатор Нирдорф лидирует с отрывом в **40 пунктов**.



...и оставляет информационный мусор на нашем пути.



От нее не скрыться.

Статистика повсюду:

в торговом центре

Эта музыка играет у нас фоном...
...потому что исследования показывают, что благодаря ей вы покупаете на 10% вещей больше!



в школе

Да, при выставлении отметок я пользуюсь графиком нормального распределения!



на кухне

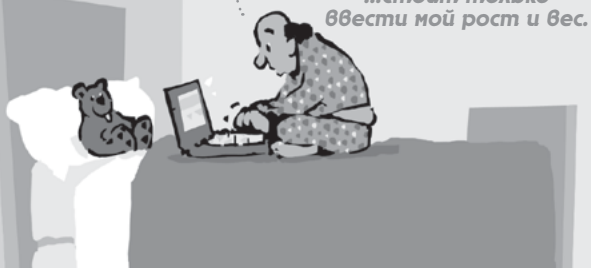
Почему я должен мыть посуду в 75% случаев?

Потому что я готовлю в 99% случаев.



в спальне

На этом сайте я смогу найти вторую половинку...
...стоит только ввести мой рост и вес.



Статистика с нами с самого рождения...

95% детей рождается на сроке между 38-й и 42-й неделей...
...так что ваши роды планируем на это же время.



...и нравится нам это или нет,
но мы и сами пополним статистику, когда умрем.

Печально.
Но, по крайней мере,
она жила дольше,
чем среднестатистическая собака.



К счастью, всему этому
есть хорошее объяснение.

**Статистика повсюду,
потому что она очень полезна.**

**Статистика помогает
предсказывать погоду...**

Существует
вероятность 95% что
завтра будет солнечно.

Но есть и шанс в 3%,
что прольется
дождь
из лягушек.



**...и систематизировать
информацию в интернете...**

Основываясь на истории
ваших покупок, я могу
составить рекомендации
для вас.

И как они
только узнали,
что я хотела
фигурку Уильяма
Шетнера*?



...и развивать медицину...

Наши исследования показали,
что при лечении рака
этот препарат на 2,5%
эффективнее плацебо,
но погрешность в расчетах
составляет 12%...

Прекрасно!
Как бы нам
его назвать?

При этом препарат
оказался отличным
слабительным!



**...и формировать модные
тенденции...**

Благодаря
статистическим данным
я понял, что джинсовые
куртки, возможно,
вернутся в моду
в этом году.

О, да твоя одежда
в стиле 1987 года,
мне нравится!

Только давай
обойдемся
без клеша.



И это еще не все.

**Статистика помогает
побеждать на выборах...**

Всего 23% моих
избирателей
считают меня
абсолютным
болваном!



**...и возводить
электростанции...**

Спровоцирует ли наша
ядерная установка
мутации у местных
жителей?



**...и зарабатывать
деньги...**

При сохранении
нынешнего
состояния рынка...

...уже завтра я буду
на 12-15% богаче!



**...и доказывать
свое
превосходство...**

Ха, на моем счету больше
хоум-ранов, чем у тебя.

Да ты использовал
стероиды, и у меня есть
статистические данные,
подтверждающие это.



Так что же делает статистику такой невероятно полезной?

Эта штука просто класс!

Тут и вилка, и нож,
и ложка, и расческа,
и соломинка...

...и отвертка,
и кусачки для ногтей,
и карандаш,
и...



Самое простое объяснение заключается в том, что статистика помогает контролировать огромное количество важных вещей...

94% всех людей,
когда-либо живших
на свете,
уже умерли...

...и 200 млн из них
умерли от чумы...

...а на дорогах
миллионы и миллионы
гибнут ежедневно...

...а вероятность того,
что в вас попадет
молния, еще выше, когда
вы играете в гольф!



...что, в свою очередь, помогает лучше понять наш сложно устроенный мир...

...и управлять им.

Исследования доказали,
что 78% людей
обожают пончики.

Поэтому, если мы
начнем раздавать
их бесплатно
на собраниях
агентов нашего
культа смерти...

...мы сможем
привлечь
новых членов!



Но настоящая сила статистики все же в другом.

Вот в чем кроется истинная причина того, почему всем нужна статистика.

Статистика помогает принимать уверенные решения...

...когда мы располагаем неполной информацией.



Позвольте объяснить, что это значит...

Представьте себе, что мы хотим узнать средний вес...

...всей рыбы в озере.

Ловись, ловись, рыбка...

Если мы узнаем, сколько в среднем весит одна рыбка...

...мы сможем понять, сколько рыбешек нам нужно ловить каждый день, чтобы спасти наших питомцев от голодной смерти!

Если бы мы осушили озеро и взвесили каждую рыбку...

...то получили бы всю необходимую информацию и высчитали средний вес.

Но по очевидным причинам мы не можем этого сделать.

По-моему, это была не лучшая идея.

С другой стороны, если мы поймаем 100 рыбешек и взвесим их...

Эти 100 рыбок весят 112 кг.

Следовательно, в среднем одна рыбешка весит 1,12 кг!

...мы получим неполную информацию о всей рыбе в озере.

Итак, теперь нам известен средний вес рыбы в этой выборке.

...но мы по-прежнему не знаем средний вес остальной рыбы в озере.

Но вот что интересно:

прибегнув к инструментарию статистики, мы можем использовать эту неполную информацию...

Статистика предполагает использование той рыбы, которую мы поймали...

...чтобы сделать доверительное суждение относительно всей рыбы в этом озере.

...чтобы судить о той, которая осталась в озере.

Правда? Как же это работает?

Наша книга как раз об этом!

Эта книга отвечает
на фундаментальный вопрос
статистики:

**КАК НАМ ИСПОЛЬЗОВАТЬ
ВЫБОРКУ...**





**...ЧТОБЫ СДЕЛАТЬ ДОВЕРИТЕЛЬНОЕ
СУЖДЕНИЕ ОБЪЕДИНЯЮЩЕЕ
О ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ?***

Все проблемы
статистики
проистекают
именно отсюда!

* Игра слов. Генеральная совокупность в англ. терминологии — population («популяция»), совокупность всех объектов, относительно которых делаются выводы при изучении конкретной проблемы. Прим. ред.

**В первой части мы научимся
делать выборку...**

...и изучать ее.

Хм, как много
рыбы.



**А затем, во второй части, мы научимся использовать выборку, чтобы
получить качественные результаты
для генеральной совокупности...**

**...используя процесс, который носит название
«статистическое заключение».**

Что же могут
эти рыбешки...

...сказать нам
о тех?



Таким образом
мы сможем обработать
большие объемы
данных...

Черт!
У нас перекос!

Это
ненормально!



И в более общем смысле мы получим
представление о том, что можно...

...высчитывать
доверительные
интервалы...

Я на 95% уверен,
что мы примерно
так же сильно
вас ненавидим.



...и проверить
гипотезы.

А я на 3% уверена,
что моя установка
по производству
ядов работает!



...и что нельзя...

Мы можем использовать
статистику, чтобы
делать доверительные
предположения...

...но их никогда
нельзя использовать
как неоспоримый
факт.

Если мы не поймем
всю рыбу...

...мы никогда не сможем узнать
со всей определенностью,
что же там творится внизу.

...делать с помощью
статистики.



**В этой книге мы сфокусируемся
на основных понятиях.**

Таких, как стандартные
отклонения...

...и распределение
выборки...

...и вероятности...

...И ДОСТОВЕРНОСТЬ!

**Если же вам интересно узнать
о деталях...**

Например, что, черт возьми,
означают все эти
формулы и символы?

**...то вы найдете их в приложении,
которое называется
«Математическая пещера».**

Часть первая

СБОР СТАТИСТИЧЕСКИХ ДАННЫХ

Не подглядывать!



Глава 1

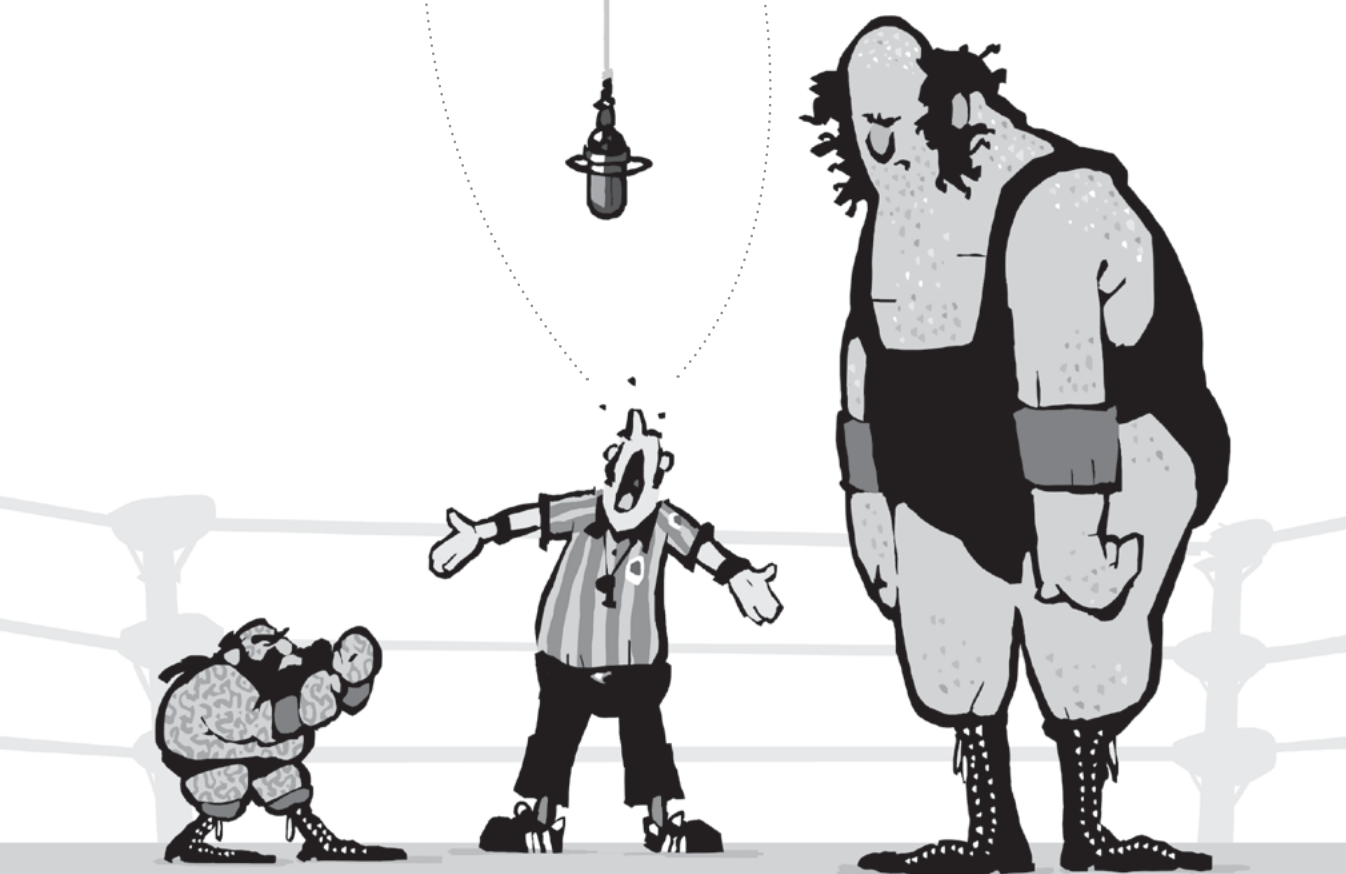
ЧИСЛА

В этом углу ринга боксер,
который весит
**50,8 триллиона
нанограммов...**

...КАРЛИК!

А в этом углу боксер,
который весит
0,193 тонны...

...ГИГАНТ!



**Как мы узнали из предисловия,
статистика — это не только
цифры.**

Доброе утро,
это оператор 3810448,
чем я могу вам помочь?



**Статистика нужна, чтобы измерить
нашу уверенность в чем-либо.**



**Как бы то ни было, статистика в действительности —
это упорная борьба с цифрами...**

С кем бы вы
хотели
сразиться?

На моем счету
147 побед
и всего
17 поражений.

А я побеждаю
в 89,6%
случаев!



....а это не всегда легко.

Эм...
Что-то я сейчас
не очень уверен
в себе.



Некоторые числа больше...

**В вашем мозгу 10^{14} нервных
соединений!**

То есть 100 000 000 000 000 соединений!



Но некоторые большие числа описывают совсем маленькие вещи...

...а некоторые маленькие.

Каждый атом в 10^{-17} раз
меньше глазного яблока.

То есмь в 100 000 000 000 000 000
раз!



...и наоборот.

**В Техасе обитает
более 500 млрд
насекомых.**



**В Солнечной системе
всего одна
звезда.**



Некоторые числа говорят о чем-то хорошем...

**Мы выиграли,
забив на два гола
больше,
чем соперники!**



Но бывает и так, что какие-то положительные числа описывают отрицательные вещи...

**Безработица
в прошлом месяце
увеличилась на 4%.**



..а некоторые — о плохом.

Индекс Доу-Джонса упал на 423 пункта! Ужас!



...и наоборот.

**Количество
убийств в городе
снизилось.**



Но и это еще не все...

Некоторые цифры выглядят пугающе...

В вашем организме
живет почти
килограмм бактерий!



...другие кажутся обнадеживающими.

Если было продано
более миллиарда
гамбургеров...

...должно быть,
они действительно
вкусные!



Некоторые говорят о серьезных достижениях...

Вирус оспы во всем
мире уничтожен
на 99,99%.



...другие не о столь серьезных...

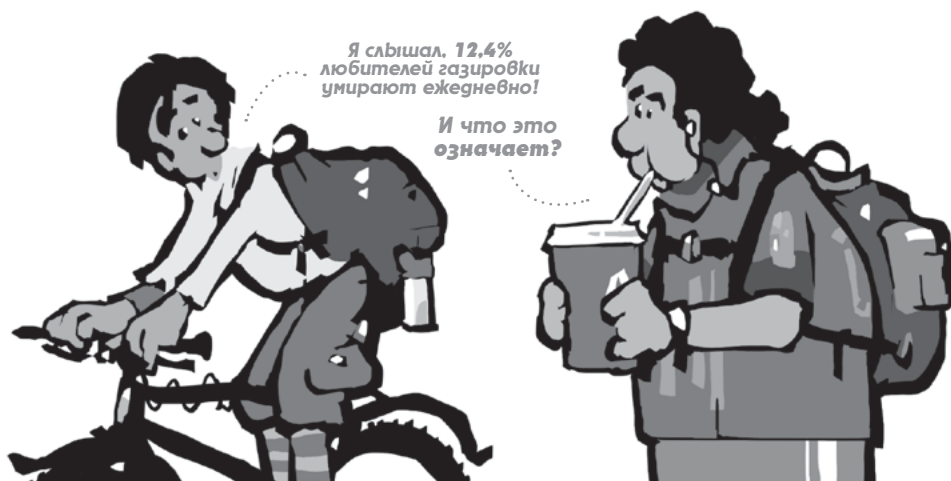
Наши вычисления
доказали,
что конец света
наступит
29 февраля
2024 года!



...и иногда сложно заметить разницу.

Я слышал, 12,4%
любителей газировки
умирают ежедневно!

И что это
означает?



**Все эти факты позволяют
с легкостью использовать числа...**

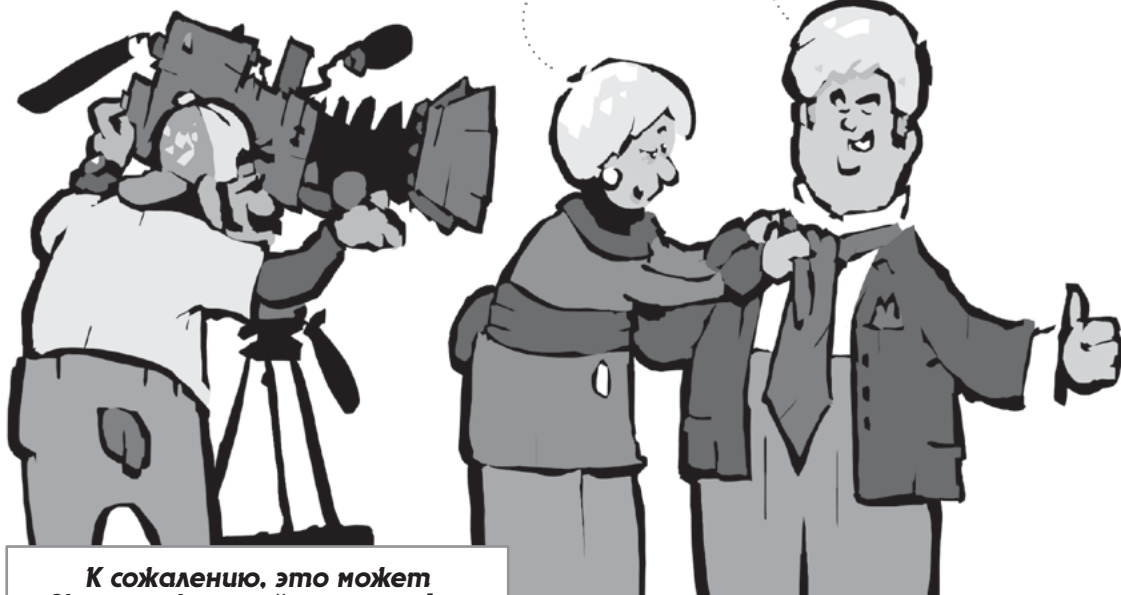
...чтобы кого-нибудь обмануть.

Если вы
наденете
этот галстук...

...все будут думать,
что вы человек
влиятельный.

А если я приведу
какие-нибудь цифры...

...все решат,
что я умный.



**К сожалению, это может
вынудить людей относиться
к цифрам с недоверием...**

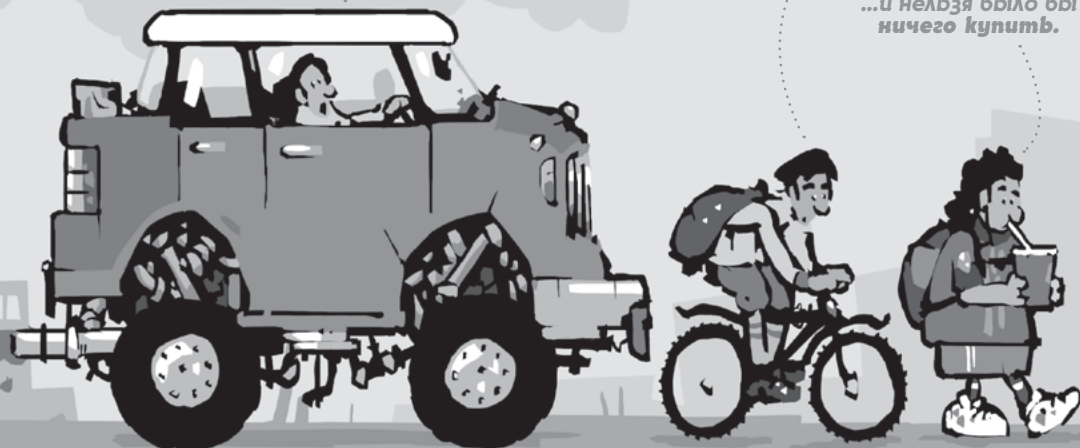
**...и не ценить их истинную
силу.**

Мне все равно, если
выброс CO_2 составит
5,5 млрд тонн...

**...это всего лишь
цифры.**

Но без них не было бы
видеоигр...

...и нельзя было бы
ничего купить.



**Решение
проблемы в том...**

Некоторые цифры
действительно
отображают
положение дел.

Но нужно помнить,
что бывает
и неверная
информация!

Как же тогда понять,
что правда,
а что ложь?



**...чтобы относиться
ко всем числам...**

...независимо
от того, большие
они...

...или маленькие...

...или и вовсе
вгоняют в сон...

Хррр...



...с долей здорового скептицизма.

Это печенье органическое на 100%
и на 98,3% подходит веганам...

...а лактоово-показатель
находится в пределах
рекомендованной нормы.

Пожалуй, возьму
печеньку,
но с долей
здорового
скептицизма.



Это первый урок нашей книги.

Помогите!

Что-то
я сомневаюсь
в этих цифрах.

И хорошо!

Пусть это ощущение
будет в радость!

В статистике
это верный подход!



Дальше мы узнаем, как инструменты статистического наблюдения помогают нам делать с помощью цифр прогнозы на будущее.

Во Вселенной по меньшей мере 100 млрд черных дыр!



Я знаю, что вы не все из них видели своими глазами...

...как же вы можете быть уверены в этом?

**Скажу пока так:
у людей всегда есть причины
жонглировать цифрами...**



Эти цифры доказывают, что я прав!

А вот эти — что ты ошибаешься!



...и будет нелишним подумать, что это за причины.

В 98% случаев это лекарство эффективно для людей с вашим заболеванием...

...которое в 14,8% случаев заканчивается летальным исходом.

Зачем они говорят мне все это?



**Не имеет значения,
уютно вам в мире цифр...**



...или нет...

**Цифра семь
приводит меня
в бешенство,
доктор!**



**...сталкиваясь с ними, вы должны
задать себе несколько вопросов.**

**Откуда они
взялись?**

**Кто приводит
эти данные?**

И зачем?



Глава 2

СЛУЧАЙНЫЕ СЫРЫЕ ДАННЫЕ

*Сколько рабочих
потребуется,
чтобы построить
мой храм?*

*И сколько пива
нам придется
им поставить,
чтобы они принялись
за работу?*



С момента сотворения мира...

...у людей есть потребность считать все, что их окружает.



Утро доброе!
Я ваш новый
господин!

Потому что он
собственноручно
задушил
764 человека!

Но
почему?

И правда, самые ранние формы письменности были придуманы, чтобы вести математические подсчеты.

Откуда мне
знать, хватит ли
рогатого скота
и зерна, чтобы
прокормить моих
людей?

...И достаточно ли
у меня воинов,
чтобы сражаться
с недругами?

Придумал!
Нужно вести учет
всего этого...

...с помощью
вот таких
небольших
зарисовок.



**По мере развития
цивилизации...**

По моим подсчетам,
твоя империя
простирается
до самого края Земли.

**...появлялось множество вещей,
которые нужно было считать.**

Отлично, тогда сколько
бычьих хвостов
нам потребуется
приготовить
для вечеринки
в честь моего
дня рождения
на следующей
неделе?



Но тут возникла новая проблема.

Назови мне точное
число врагов,
уничтоженных
каждым из моих
воинов?

Иногда невозможно подсчитать все, что мы хотим.

Ну, воинов у вас
тьма тьмущая...

...и я не могу
поговорить
с ними со всеми!

Ну, это уже
твоя проблема,
не его!

Вот поэтому когда-то давным-давно кому-то в голову пришла мысль о том, чтобы...

...исследовать выборку...

Я знаю,
что делать!
Я поговорю
с отдельными
воинами...

**...и, изучив ее, сделать выводы
о генеральной совокупности.**

...и использую
это знание,
чтобы сделать
предположения...

...обо всех
остальных!

**Использование выборки
для описания генеральной
совокупности — это умно...**

Тот факт,
что мне известно
не все...

...не означает,
что я не знаю
ничего!



**...но есть несколько нюансов,
о которых следует помнить,
прежде чем браться за дело.**



**Во-первых, руководствуясь данными выборки,
невозможно судить с абсолютной точностью
обо всей совокупности.**



Если вы хотите
знать всю правду
обо всех комарах...

...вам нужно
посчитать
и изучить всех
комаров.



**Именно поэтому статистика нужна
для того, чтобы делать максимально
точные предположения...**



**...а не быть абсолютно
уверенным.**

Давайте-ка отсчитаем
100 комаров...

...и посмотрим,
что мы узнаем
благодаря им
об остальных.



Отличный план.

Ничего нельзя
узнать наверняка,
но, по крайней мере,
вас не сожрут заживо!



**Во-вторых, если мы застопорились
на единственной выборке...**

Я могу сделать
выводы
обо всех кальмарах,
живущих в океане...

...изучив
только эти
35 особей!

**...лучше убедимся,
что мы собрали ее аккуратно!**

Э-э-э...
А мы вымыли руки,
прежде чем
трогать их?

**Потому что любая ошибка,
допущенная нами при определении
выборки...**

**...может кардинально
исказить наши
выводы о генеральной
совокупности.**

Допускается
использовать
либо дюймы,
либо сантиметры,
но не то и другое вместе.

Вы оставили
свой кофе на весах.

А это вообще
осьминог.

Чей это
тут волос?



В наши дни данные получают самыми разными способами...

Опрашивая

Подсчитывая

Взвешивая
и измеряя

Пробуя
на зуб

...и это далеко не простая работа.

**Добиваться точности измерений
бывает особенно сложно, если речь
идет о крупном...**

Возьмите-ка
линейку...

...чтобы
измерить тех
драконов...

Это помогает
сфокусироваться
на деталях.

**...или мелком
масштабе...**

...и вот этих
стрекоз.

**...или когда в процессе участвует
слишком много наблюдателей.**

Длина этой дороги
1,64 метра.

Совсем не так,
ее длина
107,9 метра!

Определение выборки также затруднительно в тех случаях, когда мы пытаемся понять, о чем думают люди...

Какой цвет более выигрышный: красный или зеленый?

Я гальтоник.

...или что они чувствуют...

Как сильно вы любите своих ближних?

Да я их ненавижу!

...или выяснить у них то, о чем они, возможно, даже не хотят говорить...

Сколько банков вы ограбили?

Зависит от того, кто интересуется.

...или когда люди преувеличивают.

Этот воин рассказывал мне, что своими руками прикончил 765 человек!

И ты ему поверил?

**Возможно, основная сложность
в формировании выборки...**

Мне предстоит
опросить
100 воинов!

Но кого именно
я должен
выбрать?



**...это понимание того, что именно следует в нее
включить.**

Кому отдать предпочтение?
Может, этим вежливым
парням, которые постоянно
сидят в кофейне...



...или тем
устрашающего вида
громилам
в спортзале?



**Следите за тем, чтобы ваше
мнение не было предвзятым...**

Если вы опросите
слишком много
вежливых
воинов...

...вам может
показаться, что
армия — гораздо более
приятное место, чем
на самом деле!

Если же вы
опросите слишком
много воинов
грозного вида...

...вы решите, что
в армии страшнее,
чем на самом деле!



В идеале хотелось бы собрать такие данные, которые бы точно отражали генеральную совокупность.



Мне бы не хотелось
выбрать 100 воинов,
которые введут меня
в заблуждение!

Это задание кажется мне просто невыполнимым...

Но как же мне понять,
насколько точно моя выборка
отражает генеральную
совокупность...

...если я даже не знаю,
как эта генеральная
совокупность
выглядит?

Твоя проблема,
не его!



...на такой случай у статистиков припасен надежный способ.

**Чтобы избежать предвзятого суждения,
мы всегда делаем случайную выборку.**

Не переживайте!

Просто
закройте глаза!

И положитесь
на волю случая!



Делать случайную выборку — это простая идея...

Давайте поместим всю армию целиком в этот шлем...



...и будем наугад вытаскивать по одному воину.

Это же как игра в «Бинго»!



...которая может на деле оказаться сложной.



ЗАБИРАЙТЕСЬ, ПАРНИ!

На практике нам нужно перебрать в уме все факторы, которые могут оказать влияние на нашу выборку...

Я не хочу включать в опрос воинов, спящих беспробудным сном в канаве...

...или тех, от которых плохо пахнет...

...или живущих в этом городе...



...и проследить за тем, чтобы это нам не помешало.

Итак, мне нужно закрыть глаза...

...и обойти всю империю...

...зажать нос...

...опрашивая воинов, с которыми меня сведет случай.



Конечно, когда мы делаем случайную выборку...



**ВСЕМ
НАДЕТЬ
ПОВЯЗКИ
НА ГЛАЗА!**



...то по-прежнему не гарантируем, что она даст нам представление о генеральной совокупности с зеркальной точностью.

На самом деле любая случайная выборка...

...будет наверняка отличаться от генеральной совокупности в целом...

...равно как и от любой другой случайной выборки.

Я отобрал этих блох совершенно произвольно.



Да эта собака вся покрыта блохами



А вот и еще 100 случайно выбранных блох.



Почему же случайная выборка так хорошо работает? Потому что мы можем взять для рассмотрения как эту выборку...

...так и любую другую...

Вот вам 100 случайно выбранных гикобразов.



А вот еще 100 случайно выбранных гикобразов.

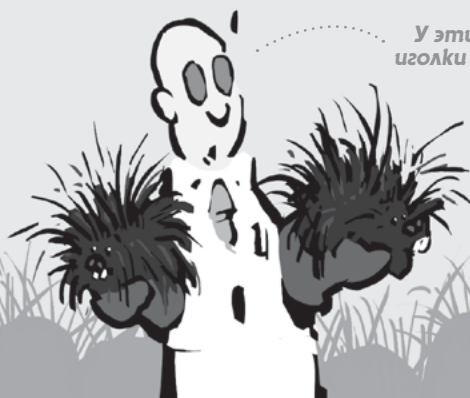


...и если уж они разные...

...это **ТОЛЬКО** по воле случая.

У этих ребят иголки глиняные...

...и это случайность.



**Делать случайную выборку
довольно сложно...**

Я провожу
исследование всей
рыбы, живущей
в море...

...и мне нужно
наугад выбрать
экземпляры
для исследования.

Остановлюсь, пожалуй,
вот на этой рыбешке,
оказавшейся здесь
в этот момент.



**...но очень важно делать
это правильно...**



Иногда стоит
дойти до края
Земли...

...или нырнуть
на морское дно!



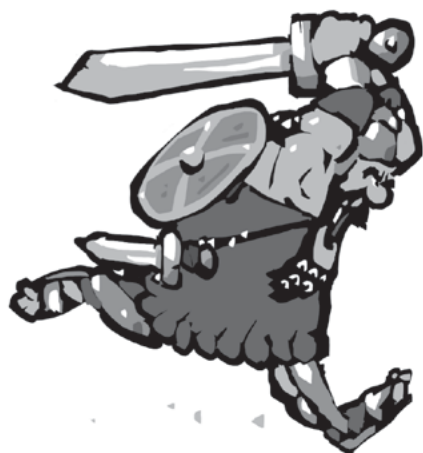
**...потому что случайная выборка* — ключ
к любому статистическому наблюдению.**

Если эти рыбешки
окажутся выбранными
не случайно...

...мы ничего
не сможем сказать
про остальных.



В этой главе мы узнали, как случайная выборка способна помочь нам избежать необъективности.



У меня сложилось бы неверное представление об армии...

...если бы я опрашивал только тех воинов, которые не собирались меня убить.



Но случайные выборки представляют собой существенную часть статистической системы, о которой мы узнаем позже.

Все инструменты, о которых мы узнаем из второй части, предполагают работу со случайной выборкой.

Кладете свою случайную выборку сюда...

...разравниваете все шероховатости...

...и получаете достоверительный интервал!



Но если ваша выборка не случайна...

...единственное, что вы получите на выходе, будет тарабарщина с кучей умных слов!



**Собранные наблюдения
называются
сырыми данными.**

Теперь все,
что от нас
требуется,
это
приготовить
их!

**Со времен сотворения мира количество сырых
данных все увеличивается...**

Как думаете,
может, у нас есть
что-нибудь еще,
на чем можно
писать?

...и увеличивается...

Извините,
но Александрийская
библиотека уже
переполнена...

...так давайте
сожжем
ее и начнем
все заново!

...и увеличивается...

У нас есть
Google!

...и увеличивается...

Теперь
у нас есть
не просто
Google...

...у нас есть
Google
в квадрате!

**...но цель, которую ставит
себе статистика, остается прежней.**

Взглянув
на случайную
выборку...

...мы сможем выдвинуть
предположения
о генеральной совокупности,
которую
она представляет.

Глава 3

РАНЖИРОВАНИЕ

Вот тут у нас
50 случайно
отобранных
носорогов...



**Мы готовы потратить время и силы
на случайную выборку...**



**...только когда нам любопытно узнать что-нибудь
о генеральной совокупности, которую она представляет.**

Итак...

Что же
вы хотите
узнать о нас?



**Иногда нам интересно узнать
что-нибудь об особенностях
опрашиваемых или составить
классификацию...**

Вы пользуетесь
дезодорантом?

Где вы
родились?

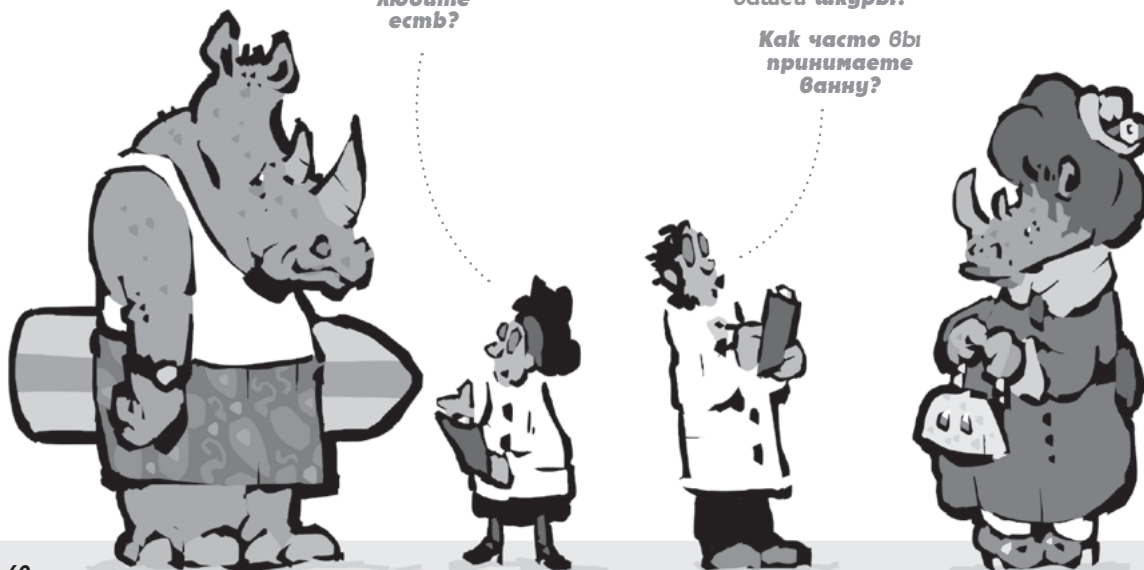
Что вы
любите
есть?

**...и иногда хочется скорее
сформулировать вопросы,
на которые мы можем
ответить с помощью
полученных цифр.**

Сколько
вы спите?

Какова толщина
вашей шкуры?

Как часто вы
принимаете
ванну?



Нам важно различать типы вопросов,
потому что от этого зависит...

Какую обувь
вы предпочитаете?

Обувь какого
размера
вы носите?

...получим ли мы
категорийные данные...

...или числовые данные...

Я предпочитаю
сапоги.

Ох, милый, я люблю
туфли-лодочки
на каблучках,
и только их.

А мне
по душе
шлепанцы.

А у меня 38,5!
У вас
37 размер.

40.

...и эти два вида данных нельзя смешивать.

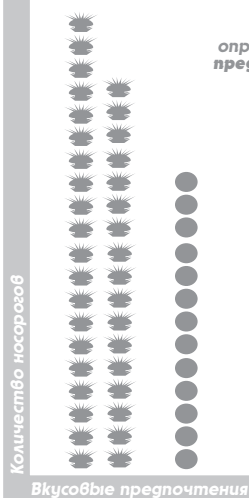
Они как
вода и масло.

Мы собираем категорийные данные...

...когда изучаем то, что можно описать только словами...



Собрав категорийные данные, мы можем сложить их стопочкой...



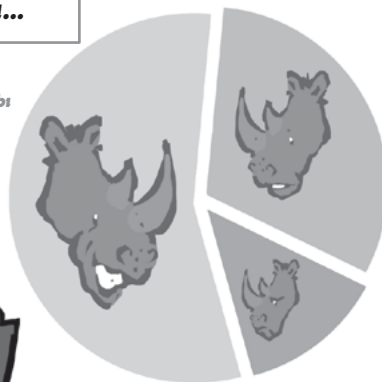
Большинство опрошенных носорогов предпочитают камням чертополох!

Хорошо бы добавить кетчупа.



...или разделить на кусочки...

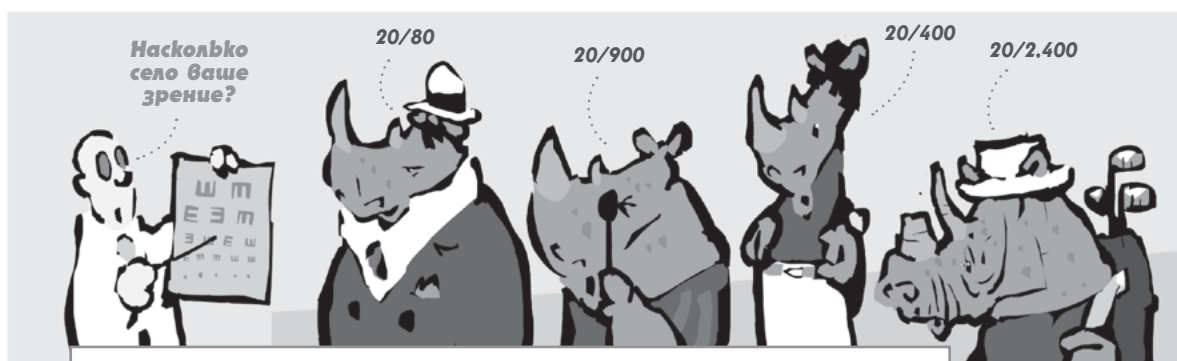
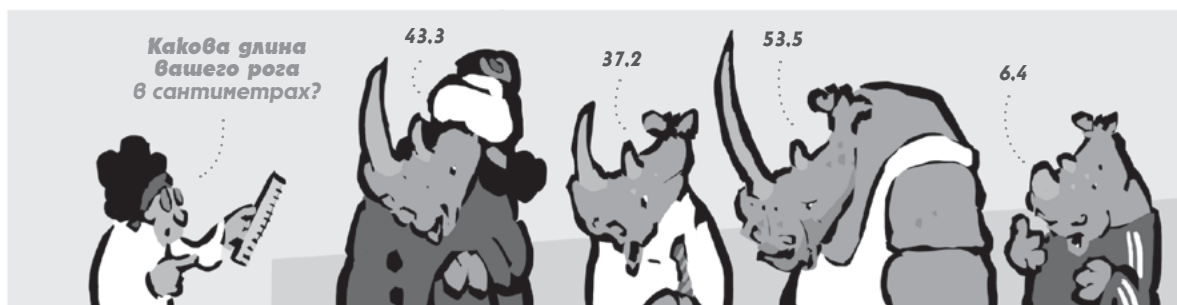
Большинство опрошенных носорогов настроены оптимистично!



...чтобы можно было получить представление о соотношении в нашей выборке.

Мы собираем числовые данные...

...когда изучаем параметры, которые можно сравнить, используя числа.



Как мы увидим во второй части книги, благодаря всем этим показателям числовые данные оказываются в целом гораздо более полезными.



Главное различие между двумя этими видами данных...

...заключается в том, что мы не можем подсчитать категориальные данные...

Какой цвет в среднем самый популярный в твоей выборке?

Ну, я бы сказал, серо-буро-малиновый в крапинку.

Не все данные создаются одинаково.

...но можем подсчитать числовые!

Какова средняя длина в твоей выборке?

0,004 метра!

Этот факт превращает в глазах статистиков числовые данные в нечто захватывающее...

Стандартное отклонение равно сигме, поделенной на квадратный корень из n !

Обожаю, когда ты говоришь со мной сухим языком чисел.

...и кажется чем-то страшным обычным людям.

Стандартное отклонение равно сигме, поделенной на квадратный корень из n !

КАРАУЛ!

**...обрушивается
вся система и требуется
перезагрузка.**

Поэтому первое, что мы делаем, собрав большое количество числовых данных...

300

**...мы рисуем картинку,
где отображаем их все.**

Даже не переживайте, если единственное, что у вас хорошо получается, это приклеивать рисунки.

Самое простое отображение числовых данных называется **гистограмма**.



Чтобы получить гистограмму нашей выборки...

...нарисуем числовую ось.



Это те числа, которые расположены на горизонтальной оси между самыми маленькими...

...и самыми большими объемами, которые мы замерили.



Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

Сверху указываем наши данные...

...показатель за показателем.

Количество носорогов 10 9 8 7 6 5 4 3 2 1

Гистограмма похожа на огромную башню из ящиков.

Каждый носорог соответствует одному ящику.

Талия у этой гамы-носорога 343 см...

Значит, ей место вот здесь...



Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

Другой вариант визуализации числовых данных представляет собой **коробчатый график/боксплот**.



Чтобы создать боксплот для нашей выборки...

...сначала снова рисуем числовую ось...



Я самый миниатюрный во всей выборке.

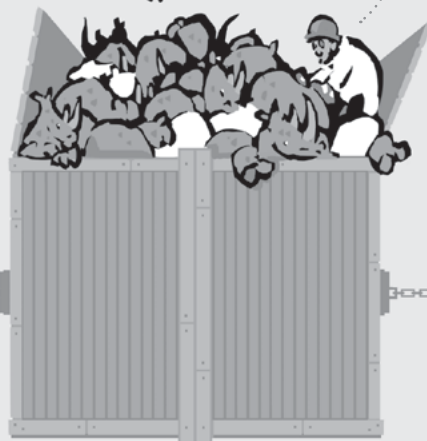
У меня самые внушительные размеры во всей выборке.



Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

...но на этот раз помещаем **промежуточные 50%** нашей выборки в один большой ящик.

Благодаря этому ящичку мы можем понять, где сосредоточена основная часть данных.



И с помощью этих планок определяем **минимальные...**

...средние...

...и **максимальные** индивидуальные значения.



Это я!



Это я!



Это я!

Обхват талии (в см) 270 280 290 300 310 320 330 340 350 360 370

Как правило, гистограммы составляют, когда нужно увидеть полную картину на основе всех наших данных...

...и выверенных деталей.

Это напоминает горный хребет.

Мы можем использовать его, чтобы исследовать вершины...

...и долины.

Вот, например, гистограмма, отображающая длину рога...

У 49 из нас длина рога колеблется от 5 до 55 см.

...а у меня 97 см!

Количество носорогов

Длина рога (в см)

...показывает, что один из носорогов намного носорожистей остальных.

С другой стороны, боксплоты могут быть особенно полезны, если необходимо сделать поверхностный обзор данных...

Боксплот — это компактная версия гистограммы.

Это все равно что смотреть на наши данные из космоса.

...или если мы хотим сравнить разные выборки или группы.

Сравнивая эту выборку...

...с той...

...мы видим, что у нижней в целом более крупные величины.

Благодаря боксплотам — «ящикам с усами» — мы быстро получаем представление о том, как данные собираются воедино...

В этой выборке данные группируются гораздо плотнее...

...чем в этой

...и понимаем, к каким выводам они нас приведут.

Странно, большая часть этих данных сдвинута влево...

...а этот огромный кусок — вправо.

**Вас может удивить тот факт,
что статистики рисуют какие-то
картинки.**

По-вашему, я могу
сделать **доверительное
суждение о генеральной
совокупности...**

**...опираясь
только
вот на эту
мазню?**

Всему, что вам
действительно
нужно знать,
вы научились
в детском саду.

**Все дело в том, что первое, что мы
всегда должны делать с собранными
данными, это просматривать их.**

**Вы удивитесь,
как часто люди
об этом забывают.**

**Потому что, хотя нас могут
привлекать более изощренные
математические инструменты...**

Эй, приятель, не интересует
непараметрический
иерархический алгоритм
Байеса?

**...именно простые картинки будут фокусировать
наше внимание на той информации, которую
на самом деле несут собранные данные.**

Прежде чем ты выйдешь
на-гора солидно
звучащие цифры...

**...нарисуй
картинку!**

Одна гистограмма
стоит тысячи
Р-значений.

Глава 4

ДЕТЕКТИВНАЯ РАБОТА

*Мило, конечно,
но что это
означает?*



**Анализировать данные все равно что
разгадывать тайну.**

Это был
профессор Плам...

...с канделябром...

...в гостиной...

...где с ним
случился приступ
гнева, когда он
пытался изучить
статистику!



**Наша главная цель — сбор улик
по одной случайной выборке...**

**...и восстановление на их основе
истории генеральной совокупности.**

Дайте-ка мне
группу случайно
выбранных
суперзлодеев...

...и я смогу
со всей уверенностью
рассказать
обо всех суперзлодеях...

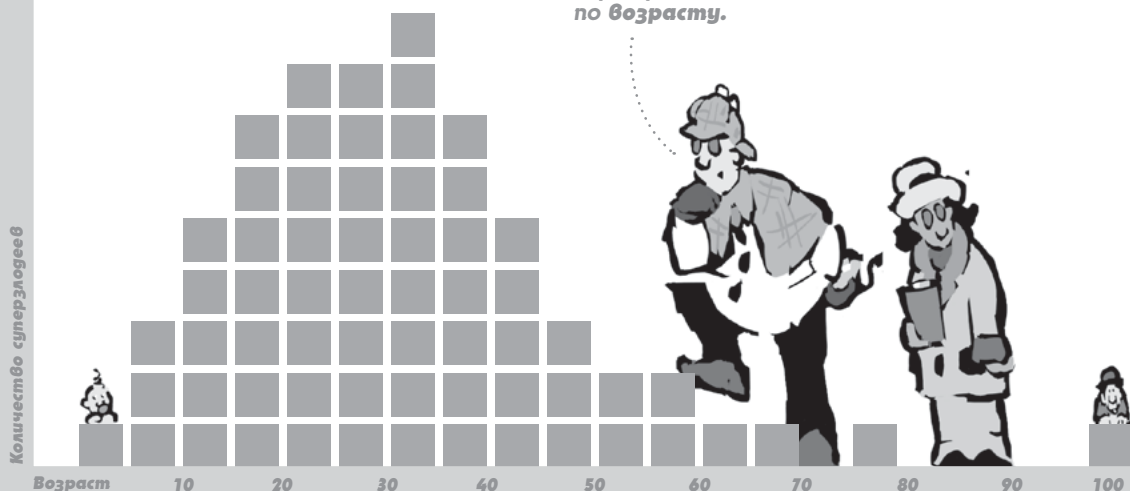
...всех-всех
суперзлодеев
в мире!



**Но первым делом нам придется
научиться выполнять самую
простую детективную работу.**

**Когда мы только приступаем к анализу
любых данных...**

В этой гистограмме
представлены 64 случайно
отобранных суперзлодея,
отсортированных
по возрасту.



**...мы всегда обращаем внимание
на четыре основные характеристики...**

ОБЪЕМ ВЫБОРКИ

Итак, сколько
у нас тут
данных?



ФОРМА

Как они
выглядят?



РАСПОЛОЖЕНИЕ

Где именно
это находится?



РАЗМАХ ВАРИАЦИЙ

Насколько
он велик?



...и мы посвятим эту главу их изучению.

Какие тайны скрывает
этот холмик?



Давай-ка поищем
какие-нибудь
погсказки.



ОБЪЕМ ВЫБОРКИ

Сколько же
здесь данных?



ОБЪЕМ ВЫБОРКИ* — первое, что нужно установить,
когда приступаешь к анализу данных...



Итак, сколько
суперзлодеев
мы отобрали
случайным
образом?



64

...и довольно просто понять, почему это так важно.

Если бы ваша
выборка состояла
из совсем небольшого
количества злодеев...

Вот нас,
например,
всего пять!

...вы бы не смогли
сделать никаких
выводов о генеральной
совокупности.

Извините, но эта ваша
картинка с данными
не очень-то мне
помогает.



Как правило, выборка большего размера оказывается полезнее!

Имея на руках всего несколько объектов исследования, мы не сможем увидеть многое...

...но если мы соберем больше случайных данных...

...наша гистограмма станет гораздо более информативной!



Как мы узнаем чуть позже, размер выборки напрямую связан с уровнем достоверности, с которой мы можем судить о генеральной совокупности.

Размер имеет значение!

Если мы добавим еще немного случайно отобранных объектов...

...мы получим большую достоверность!



К сожалению, на практике объем выборки всегда чем-нибудь ограничен.

И смотрите за тем, чтобы случайно отобранных злодеев было не слишком много, Ватсон!

Я был бы рад найти еще кого-нибудь, Холмс...

...но у нас больше не осталось наручников.



ФОРМА

Форма каждой
выборки
уникальна...

...как
отпечаток
пальца!



**Момент, когда кто-то понимает, какая форма
у выборки, может быть весьма захватывающим...**

Извините, миссис
Джонс, но ваши данные
не совсем обычные.

Осторожно!



**...потому что какой бы ни была ваша гистограмма,
она всегда имеет такую форму по какой-то причине.**

Ваша гистограмма
похожа
на верблюда...

...должно быть,
на это есть
свои скрытые
причины.



Например, мы называем распределение **равномерным**, если все исходы одинаково вероятны.



Количество ящериц

Пытаюсь запустить в жену ящерицей...

...но я неважно целюсь...



...потому есть вероятность, что его ящерица либо не долетит...

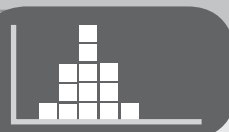
...либо просвистит надо мной...

...либо и правда попадут в цель.



Расстояние, которое пролетела каждая ящерица

Мы называем распределение данных **нормальным**, когда есть нечто превалирующее, что заставляет факты группироваться вокруг одного конкретного значения.



Количество ящериц

Я пытаюсь запустить ящерицей в мужа...

...и целюсь я гораздо лучше, чем он...



...потому вероятность того, что ее ящерица попадут в цель, гораздо выше...

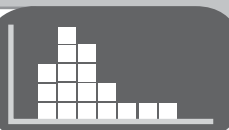
...и только некоторые, возможно, не долетят...

...или перелетят.



Расстояние, которое пролетела каждая ящерица

Мы называем распределение **смещенным**, когда по какой-либо причине в одной части находится больше данных, чем в другой.



Количество рыбешек

У меня хороший глазомер, но когда я бросаю протухшую рыбу...

...она, бывает, выскальзывается из руки при замахе.



Из-за этого многие рыбешки чаще не долетают до цели...

...чем перелетают.



Расстояние, которое пролетела каждая рыбешка

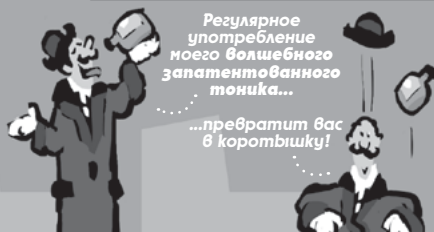
РАСПОЛОЖЕНИЕ

Где же находится
вся информация?

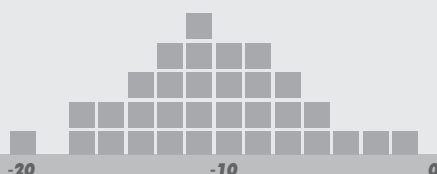


Под расположением понимают место скопления
наибольшего количества данных на оси.

Данные могут группироваться вокруг отрицательных значений...



Изменение роста
(в см)



...или маленьких
значений...



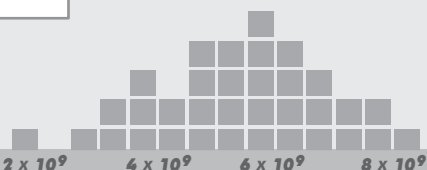
Количество глаз



...или по-настоящему
больших значений.



Возраст в годах



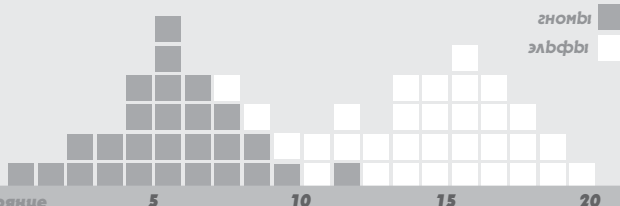
На практике статистиков часто интересует сравнение расположения
разных множеств данных.

Эльфов мы, огры,
кидаем...

...дальше,
чем гномов.



Расстояние
в метрах



Дать слову «расположение» словесное определение может оказаться делом нетривиальным...



...поэтому часто для описания информации мы используем одно значение — среднее*.

* Оно еще называется «среднее арифметическое». Чтобы узнать, как оно высчитывается, откройте стр. 214.

Получив это значение,
остановитесь:
вы у цели!



Чтобы подсчитать среднее значение, мы просто складываем все данные...

Эй, пираты, ну-ка сложите все заработанное вами за год в это ведро!

Черт!



...и потом делим их на количество данных в совокупности.

Общее количество дублонов составляет 6000...
...а пиратов у нас 50.

Таким образом, средний доход пиратов на этом корабле составляет 120 дублонов в год!



Но даже притом, что среднее значение информативно и точно как средство измерения расположения, оно не идеально.

Ух ты!
В среднем
каждый из нас
очень богат!

Тогда как же так
получается, что одноглазый
Джек не может позволить
себе искусственный глаз?



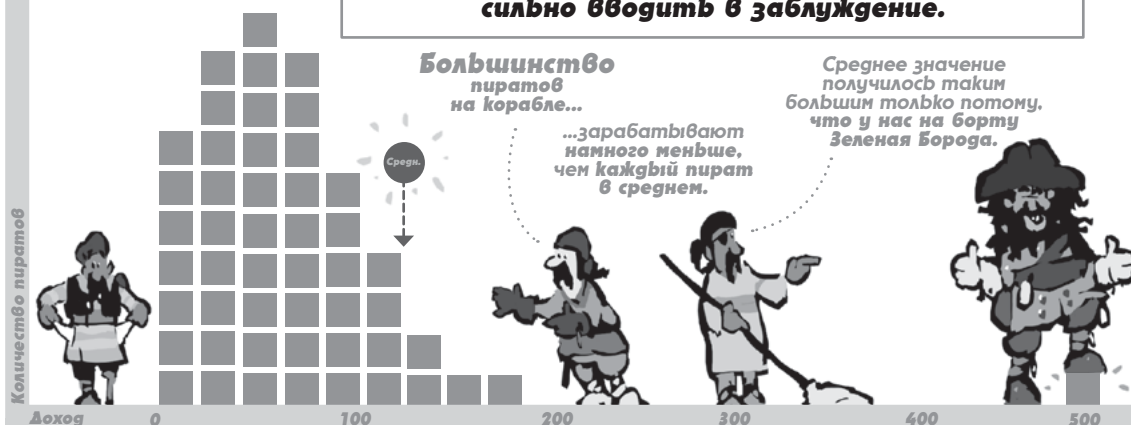
К сожалению, среднее значение может быть обманчиво.

Но тот факт, что наш средний доход составляет 120 дублонов...

...не означает, что большинство из нас настоящие богачи!

Например, если распределение смещено...

...то среднее значение может сильно вводить в заблуждение.

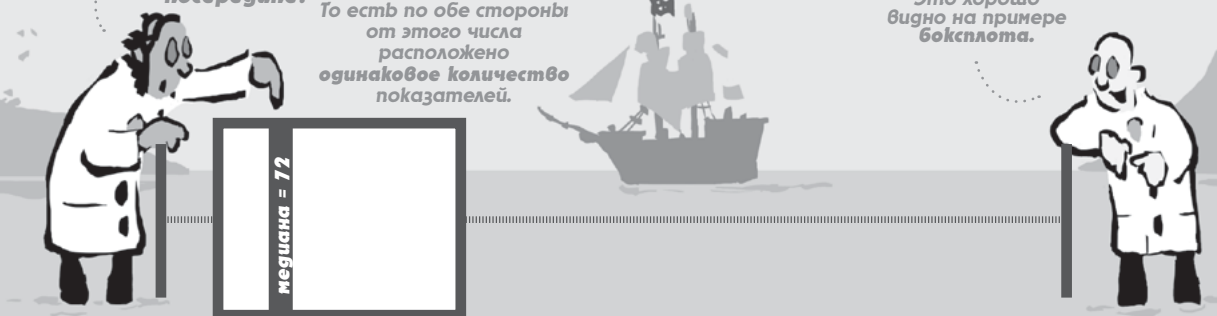


В случае смещенного распределения наиболее показательной будет медиана...

Медиана — это значение, находящееся ровно посередине!

То есть по обе стороны от этого числа расположено одинаковое количество показателей.

Это хорошо видно на примере боксplots.



...потому что благодаря этому мы лучше понимаем «типичное» значение.

Я — медиана.

Я зарабатываю 72 дублона в год.

А я представляю собой среднее значение.

Я зарабатываю 120!

Я зарабатываю 500... Йо-хо-хо!

Именно из-за меня показатели этих двоих так сильно разнятся.

Доход 0 100 200 300 400 500

Поэтому, когда некоторые с важным видом бросаются средними значениями...



Это одна из причин, по которой никогда нельзя рассматривать расположение множества данных...



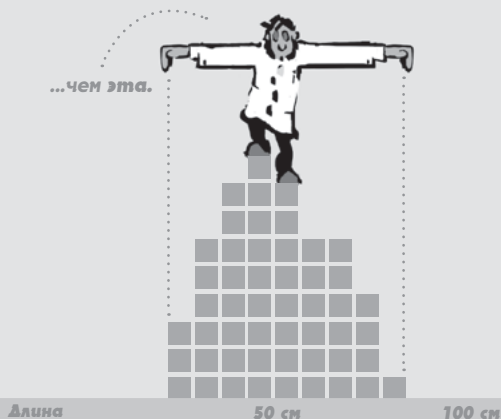
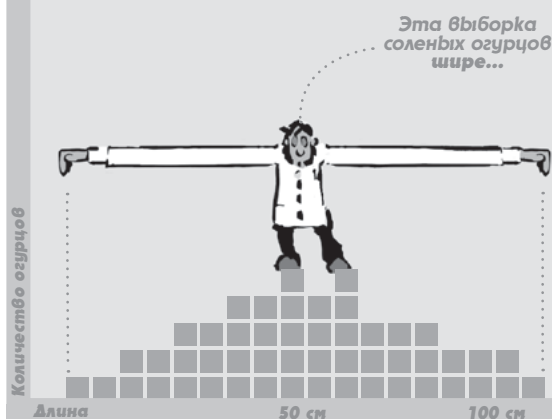
...и размах вариаций, о чем мы поговорим дальше.



РАЗМАХ ВАРИАЦИЙ



Размах вариаций — это показатель распространенности данных...



...но это также и мера разнообразия вариантов.

В этой выборке
больше вариантов...

...чем в этой.

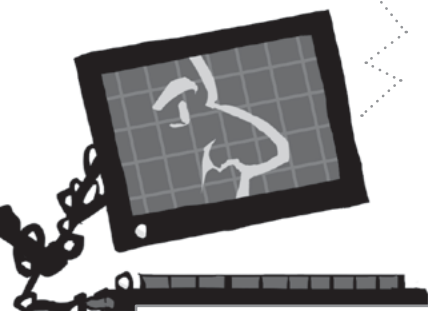


Например, если мы возьмем **выборку из 10 носов**, копированных на компьютере...



...мы не увидим **никакого разнообразия**...

Каждый нос длиной
0,23 см.



...а значит, и **размаха вариаций**.

Скукота!ща!



Количество носов

Длина 0 см 1 см 10 см 20 см

Но если мы рассмотрим **выборку из 10 носов**, нарисованных от руки...



...то увидим **большое разнообразие**...

Эти носы, нарисованные
от руки, представляют
собой линейку от 0,1 см...

...до 16,98 см.



...а следовательно, и **размах вариаций**.

Шире размах
вариаций — больше
вариантов!



Количество носов

Длина (в см) 0 1 10 20

Надежный способ измерить размах вариаций — взять весь диапазон...

...который представляет собой разницу между максимальным и минимальным показателями...

...и поделить его на четыре части.

В каждой части будет одинаковое число данных.

В этом случае 16.

Эти два «куска» в середине называются «межквартильным размахом».

Итоговое количество злодеев



Это дает нам представление о разнообразии в рамках каждой отдельной части общей выборки...

Под медианой числа идут довольно плотно...

...над медианой они имеют больший размах.



...и это особенно важно, когда мы исследуем смещенные данные.

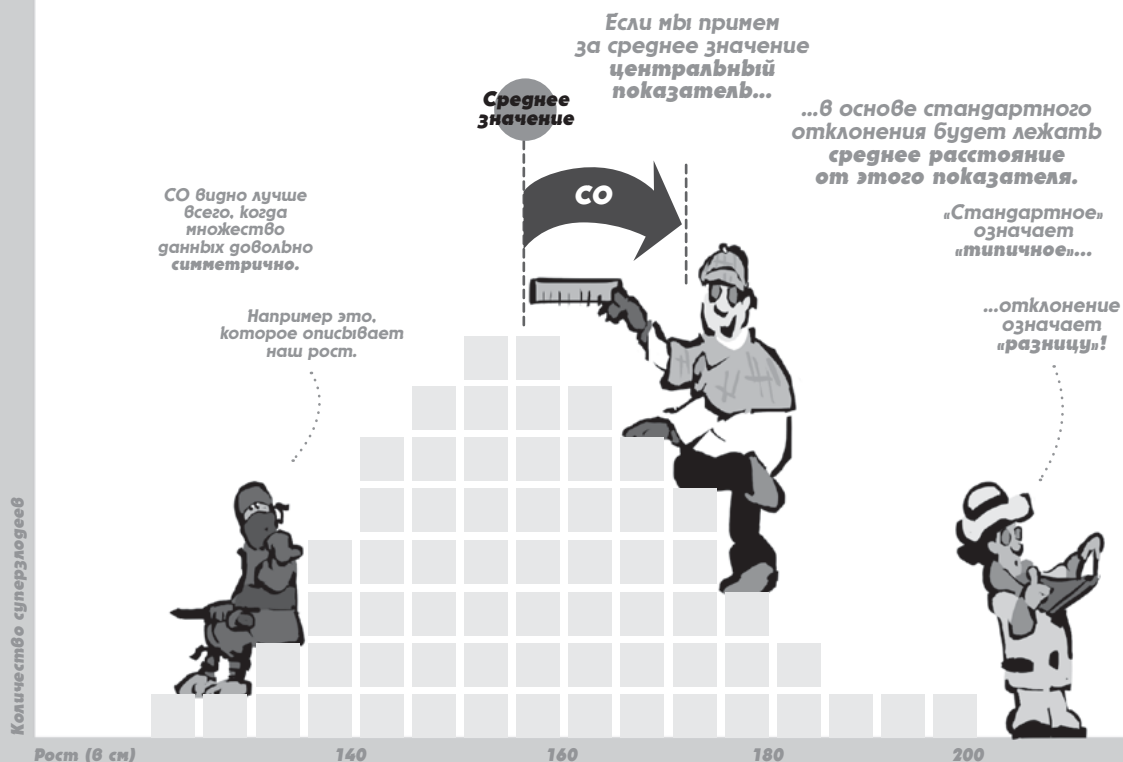
Похоже, перевес в нашей выборке злодеев на стороне стариков, Ватсон...

...вот интересно, как будет выглядеть общая картина, если Джимми Чугак сыграт в ящик?

Я не жалею.



Наиболее распространенная мера размаха вариаций — стандартное отклонение (CO)*.



К сожалению, высчитать стандартное отклонение довольно непросто.

Берем квадратный корень из всех квадратов разностей между элементами выборки и средним арифметическим значением!

Пока запомните, что у большего множества и стандартное отклонение будет больше.

ААААААА!

А чем больше стандартное отклонение...

...тем больше вариантов!

Среднее значение

стандартное отклонение



* Чтобы научиться высчитывать его, загляните на стр. 215.

В этой главе мы узнали о четырех важных характеристиках, которые изучаются в любой выборке...

Ну хорошо, вот у нас есть произвольная выборка рыбы!

Отлично, и каковы же ее **объем**...
...форма...
...расположение...
...и размах вариаций?

...вскоре мы начнем охоту за этими самыми характеристиками в генеральной совокупности.

У генеральной совокупности тоже есть **объем, форма, расположение и размах вариаций...**

...просто вы никогда не определите их со всей точностью!

Но сначала давайте применим на деле то, чему мы уже научились, чтобы разрешить спор!

Глава 5

СТРАШНЫЕ ОШИБКИ



Чаще всего, когда мы отправляемся собирать данные...



...мы хотим узнать что-нибудь важное об устройстве этого мира.

Когда эти горы выступили из моря?



Сколько людей было обезглавлено за время правления короля Генриха VIII?



Я буду нравиться девочкам, если начну носить эти штаны?



Некоторые вопросы довольно просты...

...и ответить на них можно, просто посмотрев на один набор выборочных данных.

У скольких людей в этой стране диагностирован диабет?

Давайте осмотрим 100 случайно выбранных жителей и сделаем предположение.



У вампиров плохо пахнет изо рта?

Давайте осмотрим 100 случайно отобранных вампиров и сделаем предположение.



Но другие вопросы кажутся неоднозначными...

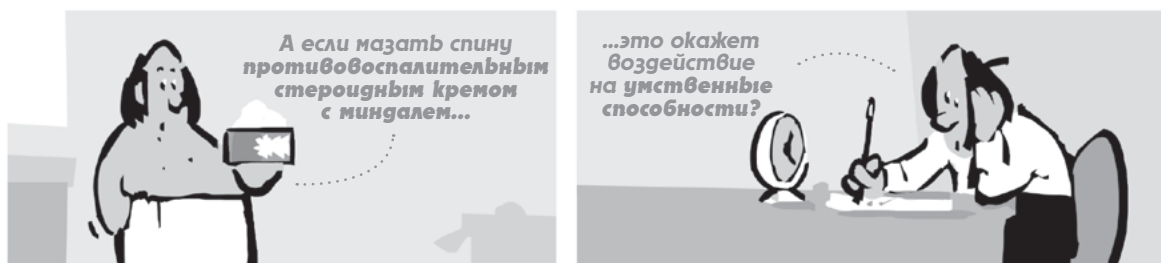
Когда они кусают больных диабетом...

...у них по-прежнему плохо пахнет изо рта?



...и требуют более комплексного анализа.

Более сложные статистические проблемы зачастую подразумевают изучение взаимосвязей...



Мы тратим много времени на то, чтобы определить, как сильно одна переменная влияет на другую...

...но помните, статистика не может быть абсолютным доказательством ни одного из наших выводов.



В этой главе
мы собираемся
исследовать взаимосвязи
двух разных
переменных...*

Если бы я был
женщиной...

...был бы я более
проворным
наездником?

...и разрешимь-таки спор.

В былые времена только викинги мужского пола
объезжали драконов.

Йууухуууу!

Но в последнее время на них стали летать
и женщины-викинги...

Йееехууууу!

...и они убеждены, что летают быстрее!

Это так и есть,
даже не спорь,
ты, шовинистская
свинья!

Чтобы понять, оказывает ли пол...



Итак, наша первая переменная.

...значительное влияние на скорость...

И наша вторая переменная.



...судьи-викинги собрали кое-какие данные.

Они сделали выборку из 50 случайно отобранных наездников...

...и 50 наездниц...

Победа за нами!

Мы выбрали их произвольно, чтобы результаты не сместились...

...из-за того что мы взяли самых быстрых наездников...

Да вы проиграете!

...или самых медленных.



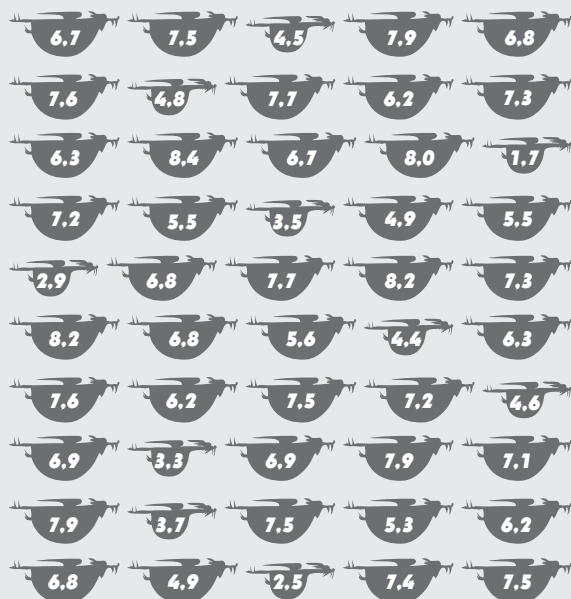
...и засеки время, за которое они преодолеют километр.

Пусть победят самые быстрые наездники!



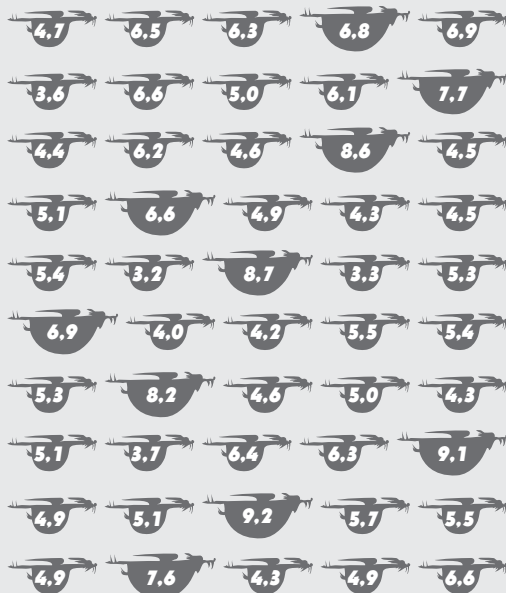


Вот результаты 50 выбранных наугад мужчин-наездников...



Время
в секундах.

...и 50 случайно отобранных наездниц.



Из этого набора сырых данных мы легко можем высчитать два средних значения...



Сложите все показатели мужчин...

...и разделите на 50.

В среднем наездникам понадобилось 6,3 секунды.

Сложите все показатели женщин...

...и разделите на 50.

В среднем наездницам понадобилось 5,6 секунды.



...и сравнить их.

В среднем наездницы оказались быстрее!

Чуть позже мы узнаем, что эти выборочные данные говорят обо всей изучаемой совокупности...

...а пока сконцентрируемся на самих выборочных данных.



**Благодаря этому сравнению
конфликт, кажется, исчерпан.**

Наше выборочное
среднее значение
выше вашего!



Но мы пока разобрали только одну составляющую общей картины.

**Остерегайтесь
поспешно найденных
средних показателей!**

Помимо этого, нужно
еще смотреть
на форму,
расположение
и размах вариаций.



**Чтобы получить более точное представление
о данных, нам необходимо нарисовать картинки.**

Подойдите
поближе и давайте
посмотрим, о чем нам
говорят эти цифры.



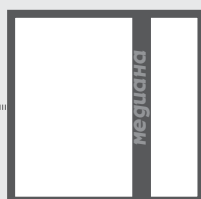
Это было вполне предсказуемо,
что картинки получатся разные...

Обе группы
смещены?

...если мы сравним боксплот
с показателями мужчин...

Так вышло, что
в целом мы преодолели
расстояние медленнее...

...зато у нас
больше вариаций
по скорости, которые
видны на более
быстрой
стороне!

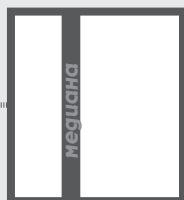


Секунды 1 2 3 4 5 6 7 8 9 10

...с аналогичным исследованием
женских показателей.

Мы затратили меньше
времени на прохождение
дистанции...

...но у нас получилось
больше вариаций
на более медленной
стороне!



Секунды 1 2 3 4 5 6 7 8 9 10

Но почему же
гистограммы обеих
групп смещены...

...в разные стороны?

Предполагаю,
что это
проделки
злых сил!



И мистика только усиливается...

Кажется, у обеих групп по две высших точки!

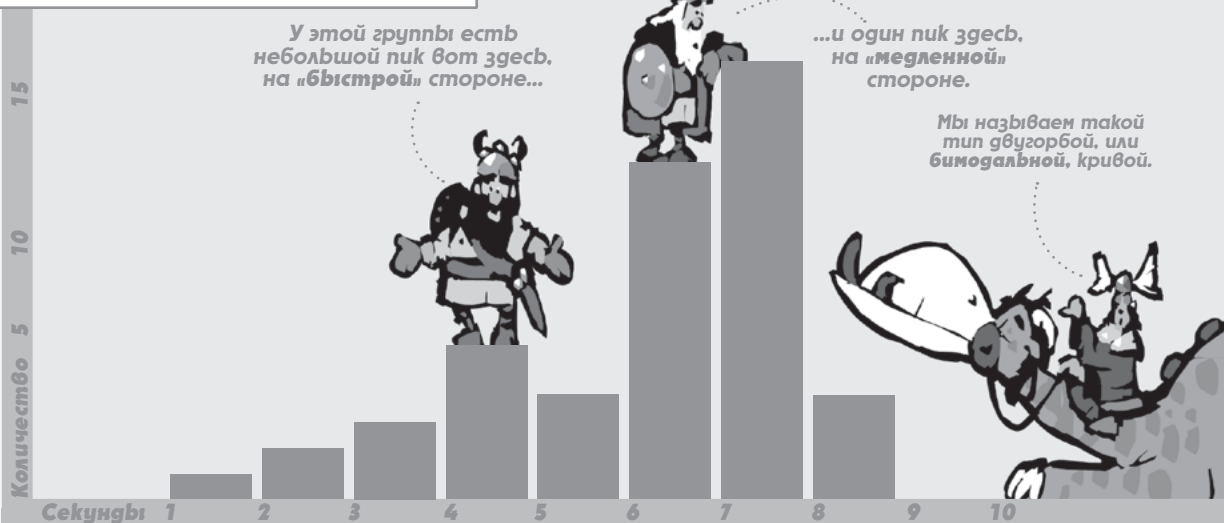


....если посмотреть на гистограммы с мужскими показателями...

У этой группы есть небольшой пик вот здесь, на «быстрой» стороне...

...и один пик здесь, на «медленной» стороне.

Мы называем такой тип двугорбой, или бимодальной, кривой.

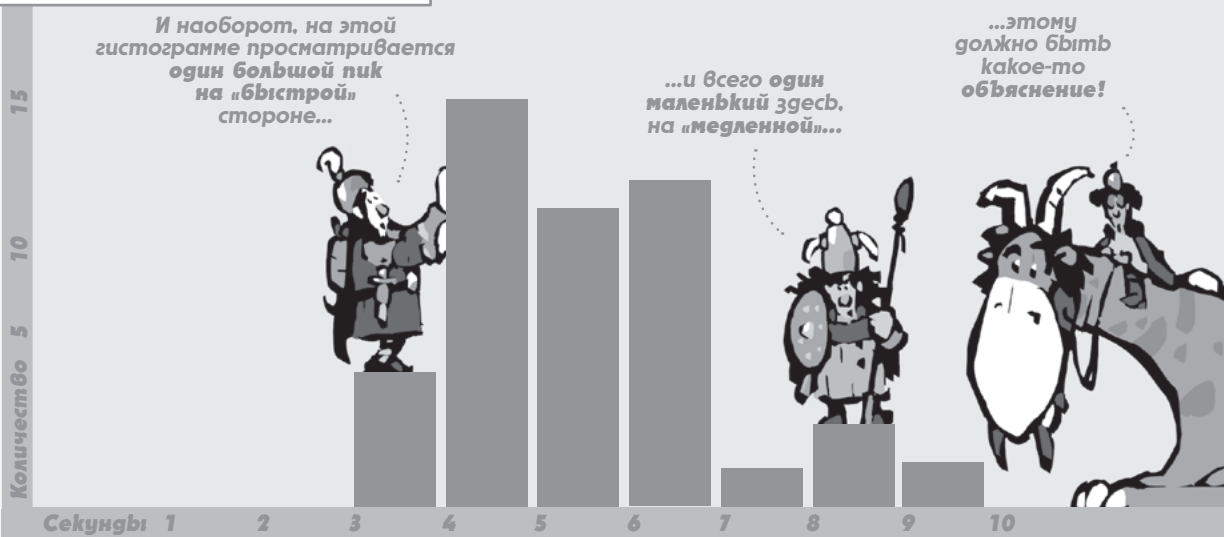


...и на женские показатели.

И наоборот, на этой гистограмме просматривается один большой пик на «быстрой» стороне...

...и всего один маленький здесь, на «медленной»...

...этому должно быть какое-то объяснение!



Это только доказывает, что соотношение двух наших переменных может на деле оказаться не таким простым, как мы думали!

Если тот факт, что ты женщина,

...заставляет тебя летать быстрее...

...тогда почему на обеих гистограммах есть смещения и присутствует мистическая двугорбость?

Помните, какими бы ни получились гистограммы по вашим данным, на то всегда есть причина.



Теперь основная задача в том, чтобы понять, почему данные выглядят именно так...

Смещения и двойные пики?

Сдается мне, мы упускаем что-то важное.

...мы можем выяснить это, поискав другие переменные, которые могут оказывать влияние.

Что же еще может сказываться на скорости наездников?

Может, что-то связанное с дистанцией?

Может, конечно, но я сомневаюсь...

...потому что это что-то влияет одинаково на обе группы наездников.

Может, все дело в том, сколько наездники весят...

...или во что они одеты?

Может, конечно, но я что-то сомневаюсь.

Помнишь, мы же выбирали их наугад.

Выясняется, что пока мы концентрируемся на половой принадлежности участников и скорости...

...мы совершенно забываем о драконах!

Эй, мы и есть ваша третья переменная!

Дело в том, что драконы бывают двух видов...



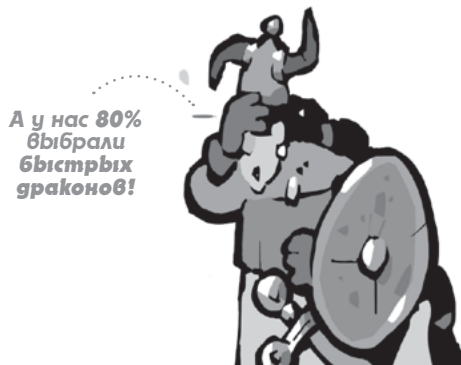
...и мужчины-наездники, как правило, предпочитают драконов покрупнее, которые оказываются менее расторопными...



...в то время как дамы отдают предпочтение драконам поменьше, но пошустрее!



Получается, неудивительно, что наездники в целом оказались быстрее.



Если мы примем во внимание тот факт, что наездники разного пола предпочитают неодинаковых драконов...

Ну да, я люблю драконов побольше, ты что-то имеешь против?

...то при подсчете среднего времени для обоих типов наездников и драконов...

...мы получим весьма неожиданные результаты.

Мы изучили сырые данные и нашли средние показатели этих характеристик.

	миниатюрные драконы	крупные драконы
наездники	3,6	6,9
наездницы	5,1	7,9

Мужчины-наездники оказались быстрее независимо от типа дракона!

Выходит, что наше первое заключение...

...оказалось не просто обманчивым...

Мы думали, что наездницы были быстрее...

...а на самом деле мы просто выбрали более быстрых драконов!

...а в корне неверным!

В целом вы, конечно, можете быть быстрее...

...но если мы примем во внимание, что драконы бывают совершенно разными...

...окажется, что быстрее мужчины!



**Пока мы были заняты изучением
связи двух переменных...**

А пол
наездника...

...влияет
на скорость?

**...мы совсем забыли
о возможной третьей
переменной, которая
все это время была
где-то поблизости...**

**...и которая в результате сделала несостоятельными
все наши заключения.**

Мой совет:
не забывайте
про драконов.

**К сожалению, скрытые переменные могут внести
неразбериху в любой статистический анализ...**

**Во Вселенной
полно
переменных!**

О некоторых
нам известно...

**...но есть и такие,
о которых мы
не догадываемся.**

**...и одна из обязанностей статистика
как раз и заключается в поиске таких переменных.**

**Мораль этой истории
заключается в том, что....**

**Будьте
бдительны
и помните
про скрытые
переменные!**

**...всякий раз, когда
нам кажется, что мы
нащупали связь между
двумя переменными...**

**Будешь есть
только капусту...**

**...увеличишь
продолжительность
жизни!**



**...может случиться, что есть и другие
факторы, оказывающие влияние
на наши заключения...**

**...Если только
при этом не
будешь забывать
и о регулярных
занятиях
спортом.**



**Если пират будет
получать даже
на несколько
дублонов
меньше...**

**...его это
сильно
разозлит!**



**...Если только дело
не в нелепости
обеда, который
на самом деле
может испортить
настроение!**



Детей...

**...приносят
аисты.**



**...Если
только они
не приходят
в этот
мир иными
способами.**



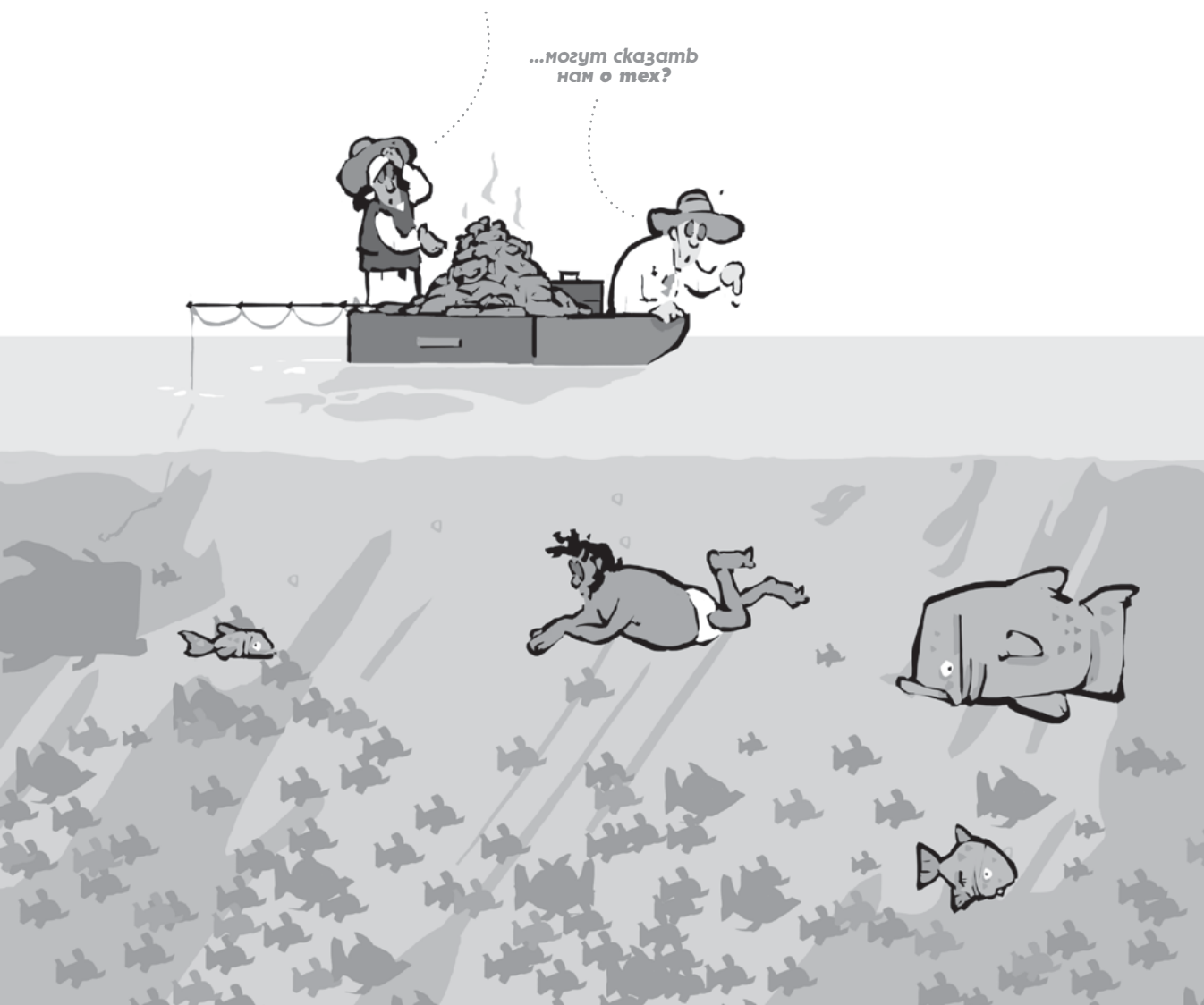
**...и если мы их не найдем, мы рискуем поверить
в то, что на деле не будет правдой!**

Глава 6

ОТ ВЫБОРКИ К ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Итак,
что же
эти рыбешки...

...могут сказать
нам о тех?



**Пока что мы говорили в основном
о выборах.**

Вот у нас тут есть
50 рыбешек, выбранных
наугад и распределенных
по весу!

**Но помните, наша
конечная цель —
использование
выборки...**

**...на которую
мы можем
посмотреть!**

**...для получения объективных
выводов о генеральной совокупности.**

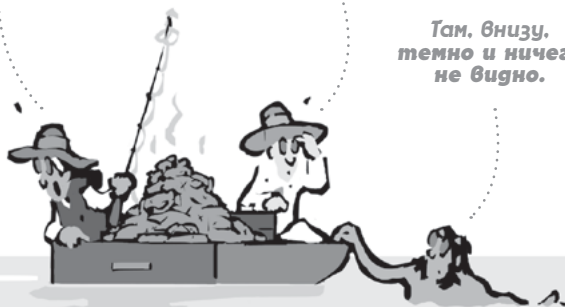
**...на которую
мы никогда
не сможем
посмотреть!**

И это создает проблему:

**как мы можем быть
уверены в информации
о генеральной
совокупности...**

**...если никогда
не сможем
посмотреть
на нее?**

**Там, внизу,
темно и ничего
не видно.**



**Во второй части нашей книги
мы поставим этот вопрос ребром...**

**Мы узнаем
о статистическом
предположении!**



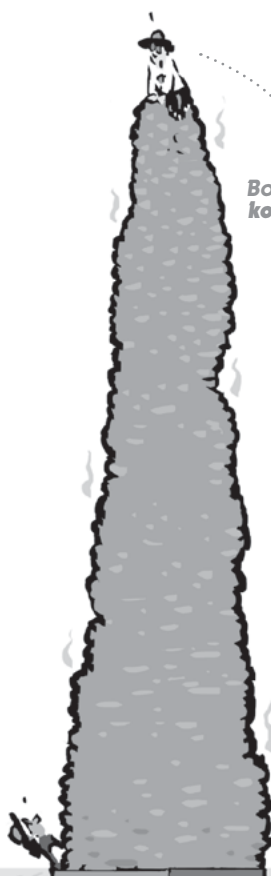
**...но прежде чем мы начнем, давайте проясним некоторые
ключевые термины, которые будем использовать.**

**Мы уже знаем,
что упорядоченные данные
нашей выборки в виде графика...**

Вот 50 случайных
рыбешек,
которых
мы только что
поймали.



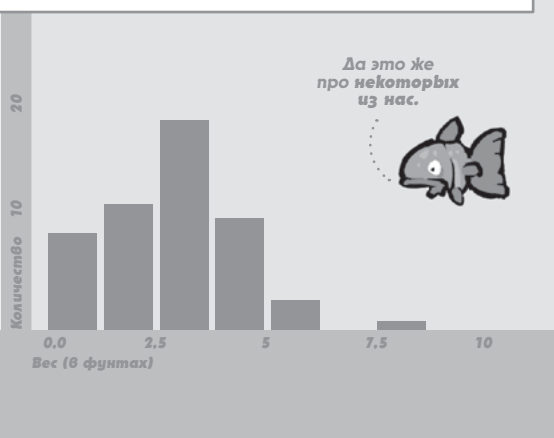
**Но если бы нам удалось собрать
вместе данные обо всей
генеральной совокупности...**



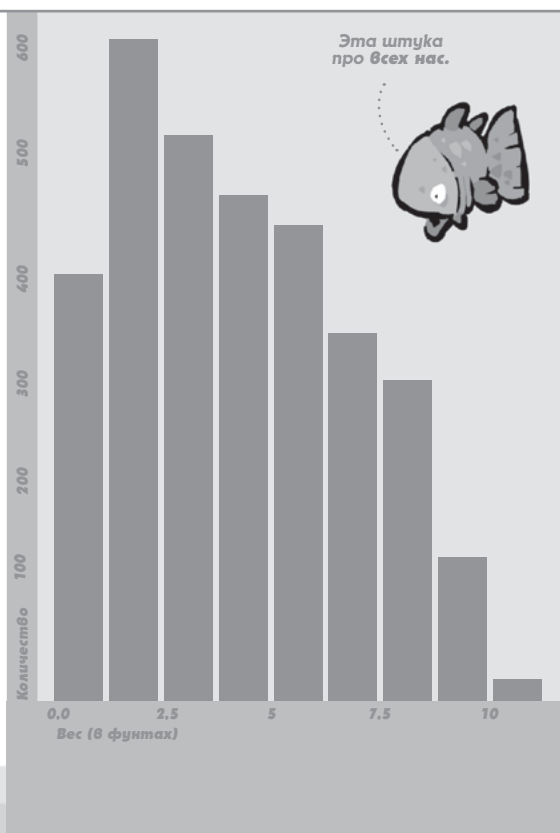
Вот вся рыба,
которая есть
в озере!

Запомните,
в реальности вам
никогда не удастся
увидеть целиком
все совокупное
распределение...

...называются гистограммой.



**...мы бы назвали получившийся результат
распределением генеральной
совокупности*.**



...если бы это было
в ваших силах,
вы бы спокойно
обошлись
без статистики.



* См. стр. 216.

Мы уже знаем, что у выборки на гистограмме есть определенные важные показатели...

У той кучи рыбы, которую мы поймали...

...есть форма...

...расположение...

...и размах вариаций...

...и мы знаем их!



Среднее значение = 3,7



$CO = 1,9$



...а тут вдруг оказывается, что у совокупного распределения тоже есть эти показатели.

У всей совокупности рыбы в озере...

...тоже есть своя форма, расположение и размах вариаций...

...но мы никогда не узнаем всех этих параметров с абсолютной точностью.



средняя величина = ?



$CO = ?$



Чтобы различить их между собой, мы называем показатели в выборке «статистическими величинами»...

Например, наше выборочное среднее значение представляет собой статистическую величину...

...и такую же величину представляет стандартное отклонение в выборке.



...а показатели в совокупности — «параметрами»*.

Например, наш общий средний показатель по совокупности — это параметр...

...точно так же, как стандартное отклонение в генеральной совокупности.



* См. стр. 216.

Иными словами, единственное, что заставляет нас отправляться собирать статистические данные...

Нам известно, что средний вес рыбы в этой случайной выборке составляет 1,68 кг...

...это наше любопытство: нам интересно, какими будут параметры.

...но что нам на самом деле важно, так это справедлива ли эта цифра для всей рыбы в озере.

Статистические данные — это то, что мы, собственно, подсчитываем и о чем можем судить с всей определенностью...

...а параметры — это то, что мы бы хотели знать, но о чем можем только строить предположения.

Статистические данные — те цифры, на которые мы смотрим.

Параметры — цифры, которые мы ищем.

Пусть мы и никогда не сможем посмотреть на параметры своими глазами...

...но, к счастью, у нас есть статистические данные, чтобы определить параметры.

Прихвати с собой статистические данные...

...мы отправляемся на ответственное задание!



На самом деле статистические данные
помогают найти **самые разные виды**
параметров.



Мы подробно остановимся на **каждом отдельно.**

Мы будем учиться
использовать статистические
данные, которые находим
в случайной выборке...

...чтобы определить **средние**
значения в совокупности,
которую она представляет.

Объем выборки?

— Есть!

Среднее значение
выборки?

— Есть!

Стандартное
отклонение
в выборке?

— Есть!

Итак,
мы готовы,
пойдем искать
параметры!

Они должны
быть где-то
тут!



**Как нам уже известно, мы никогда
не сможем использовать
статистические данные...**

**...чтобы определить параметры
с точностью.**

Я могу изучить
50 случайно выбранных
пришельцев, которых
я привез из другой
галактики.

Но и в этой галактике есть еще
множество обитателей, которых
мы никогда не сможем изучить.

**К счастью, статистики придумали способ,
как связать одно с другим...**

Ого! Мы можем
использовать
эту форму как луну...

...чтобы
поподробнее
рассмотреть
генеральную
совокупность!

...в следующей главе мы будем говорить как раз об этом!

Часть вторая

ПОИСК ПАРАМЕТРОВ



Глава 7

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА

Эта глава
о великом
открытии...

...благодаря которому
все, что есть
в оставшейся части
книги, становится
возможным...

...и оно имеет отношение
к средним значениям.



Давайте представим себе,
что нам нужно узнать среднее
значение в определенной
совокупности.

Сколько газировки
американец
выпивает в день?

Вкусно!

Класс!

А теперь представьте, что мы идем и делаем множество
случайных независимых выборок из генеральной совокупности.

У нас есть
50 выбранных
наугад
американцев.

Тут еще
50 выбранных
наугад
американцев.

И еще одна
группа
из 50 выбранных
наугад
американцев.

В каждой выборке
50 выбранных
наугад
американцев.

Мы складываем
каждую выборку
в мешок, чтобы было
проще следить за ними.

Эй, чур,
не толкаться!

Оказывается, если мы высчитаем среднее значение в каждой выборке...

Например, среднее значение в нашей выборке — 487 миллилитров.

А в нашей — 366 миллилитров.

А тут 522 миллилитра.

186.

452

186

522

...а потом расставим их по порядку и разместим одна на другой...

У нас получится гистограмма со средними значениями.

366

522

600

Уф-ф...

Средний показатель ежедневного потребления газировки, в мл

300

450

600

750

...все это множество средних значений в конце концов группируется!

Мы готовы к тому, что могут получиться экстремальные средние значения, типа такого вот.

Но большинство средних величин скапливаются вокруг этого показателя.

От 425 до 600 миллилитров в день.

Хм-м-м.

186

497

515

586

376

452

565

366

509

522

600

654

Средний показатель ежедневного потребления газировки, в мл

300

450

600

750

И это еще не все.

Оказывается, что чем больше
выборочных средних значений
вы собираете вместе...

Принесите еще!

Нам нужно
еще сто тысяч
миллионов
данных!

...тем более явно выраженное
нормальное распределение
приобретает их множество.

Помните о том,
что каждый мешок —
это отдельная выборка...

...и мы их распределяем
в зависимости
от среднего значения
в каждом мешке.

Это большое
открытие!

Средний показатель
ежедневного потребления
газированной, в мл

300

450

600

750

Это нормальное распределение имеет
определенное математическое выражение*.

Но пока запомните, что нормальное
распределение имеет колоколообразную
симметричную форму.

На самом деле
она выглядит
именно так!

$$h_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



* См. стр. 217, там об этом написано более подробно.

**И эта формула подходит
для вычисления среднего значения
выборки из любой совокупности.**

Вот пирамида, составленная
из разных видов чешуй драконов...

...отобранных
в случайном порядке
и распределенных
по среднему весу.

А вот случайные
выборки лап
ящериц...

...отобранных
по средней
длине.

И неважно, какой формы сама по себе совокупность.

Она может быть
такой...

...или такой!

Равномерная форма,
смещенная, обычная,
ненормальная —
да какая разница!

**В конце концов, чем больше средних значений
вы соберете, тем более нормально-распределенную
форму получите.**

Куполообразная
и симметричная
форма!

Очень плавно
нисходящая
с обеих сторон.

Это самая
красивая форма
в статистике!

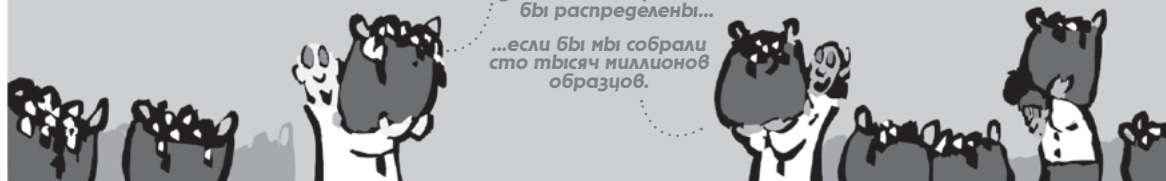
**Формально вот такое
огромное множество
представляет собой уже вид
распределения выборки*.**

Так статистические
данные выборки были
бы распределены...

...если бы мы собрали
сто тысяч миллионов
образцов.

Частота

Средние значения



* Определение см. на стр. 217.

Ну и небольшой
приятный бонус:

Эта форма
самая красивая
в статистике...

...и ей нравится слушать
хеви-метал?!


**оказывается, что центральный показатель
в огромном множестве средних значений...**

Это среднее
значение
всех средних
значений!

Принесите-
ка нам еще
выборки!

Но это работает,
только когда
у нас огромное
множество
выборок!

**...равен центральному показателю генеральной
совокупности, которую представляет выборка.**

Среднее
значение
всех средних
значений...

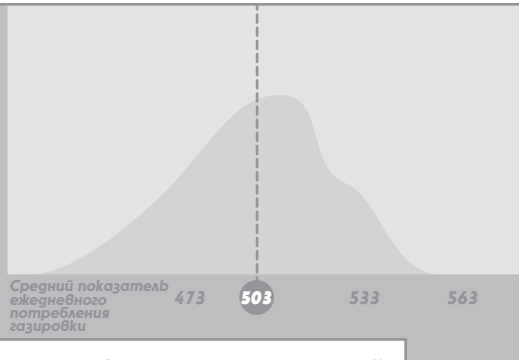
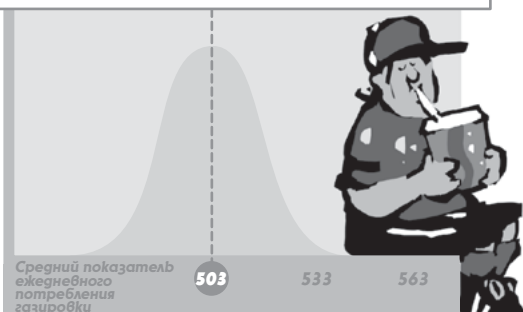
равняется
среднему значению
генеральной
совокупности.

Генеральная
совокупность
может иметь
и такую форму...

...мы никогда
не можем знать
наверняка.

Например, если это множество средних значений в выборке, составленной по количеству газировки, выпиваемой ежедневно, будет центрировано по отметке **503** миллилитра в день...

...то генеральная совокупность будет центрироваться по этому же показателю!



Это происходит потому, что огромное множество средних значений гарантированно будет иметь симметричную форму.

В конце концов для **каждого** среднего значения выборки, получаемого с помощью показателя, который **ниже** среднего значения генеральной совокупности...



Нормальное распределение всегда симметрично.

...мы гарантированно получим **другое** среднее значение выборки с помощью показателя, который **выше** среднего значения совокупности.

Конкретно эти 50 случайным образом отобранных американцев пьют **немного** газировки.



А эти 50 случайным образом отобранных американцев пьют **очень много** газировки.



А вот и еще один приятный бонус:

Я умер и попал в рай!



оказывается, что огромное множество средних значений...



Не забудь, в этом множестве сто тысяч миллионов выборов.



...как правило, тоже будет уже, чем генеральная совокупность, которую оно представляет.

Иными словами, множество средних значений имеет меньший размах вариаций...

...что означает, что будет меньше самих вариаций!



А вот насколько уже, будет зависеть от размера каждой выборки.

Если мы увеличим размер выборки, то самое большое множество будет выглядеть скорее не так...

Короткий и широкий холмик.



...а вот так.

Длинный и узкий горный пик.



Обратите внимание, что обе величины имеют нормальное распределение.

Но у того, что поуже, стандартное отклонение меньше.



Можно включить интуицию и понять, почему больший размер выборки дает более узкое множество средних значений.



Если в каждой выборке только один американец...

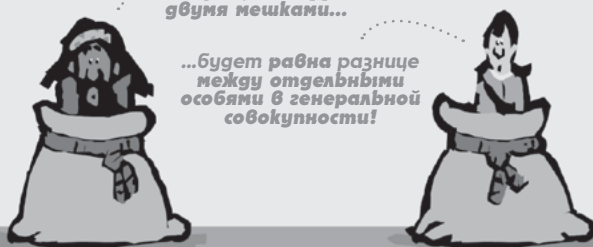
...то размах выборки столбца средних значений будет ровно таким же, как размах выборки генеральной совокупности.

Один мешок — одна выборка.



Разница между двумя мешками...

...будет равна разнице между отдельными особями в генеральной совокупности!



Но если уместить в каждой выборке всех американцев из генеральной совокупности...

.... ТОГДА распределение столбца средних значений будет равно **НУЛЮ**.

Стандартное отклонение в громадном множестве...



Между мешками не будет вообще никакой разницы!

Уф!



В любом случае, математическое соотношение — понятие точное.

Стандартное отклонение в громадном множестве...

...равняется стандартному отклонению генеральной совокупности...

...поделенному на квадратный корень объема выборки!



Ужа-ас!



Итак, к чему же сводится
наше великое открытие:

получается, что
огромное множество средних значений
случайной выборки стремится
К НОРМАЛЬНОМУ РАСПРЕДЕЛЕНИЮ!

Помните,
все выборки
одного и того же
объема.

Они все из одной
и той же генеральной
совокупности.

И их сотни тысяч
миллионов!

Ты выборочное
распределение
моих желаний.

Только
посмотрите
на эти кривые.
Как они красивы!

Цветочные корзины сортированы по среднему размеру

Все они центрированы по среднему
показателю генеральной совокупности...

...но их распределение
уже, чем у генеральной
совокупности.

И неважно,
как распределены
выборки...

...и какова
форма...

...или
генеральная
совокупность!

Официально мы называем
это открытие
**центральной
предельной
теоремой (ЦПТ)*.**

Было бы здорово,
если бы оно
имело было
более поэтичное
название.



* Откройте
стр. 217–218,
чтобы узнать о ЦПТ
подробнее.

За долгие годы статистики выработали формулы,
которые объясняют, почему ЦПТ работает.



Она работает только в том случае,
если каждая выборка будет случайной...



...а также при условии, что выборка
достаточно большая.



А вот что ЦПТ представляет собой в математических терминах:

Мы, конечно, можем
ожидать, что
огромные множества
средних значений
выборки окажутся
стандартными...

...и будут центрированы
по среднему значению
генеральной
совокупности...

...со стандартным
отклонением,
равным...

...стандартному
отклонению генеральной
совокупности,
поделенному на квадратный
корень объема выборки².

Вот это да!

1. Но только если выборки делаются случайным образом и размер каждой достаточно велик
(больше 30 или около того).

2. Для любителей математики: обратите внимание, что весь прямоугольник будет уже,
если размер выборки больше.

Не потеряй эту
цианотипию,
потому что мы будем
использовать ее позже.

**Но вот способ попроще,
как все это запомнить:**

**средние значения
случайной выборки
стремятся к среднему
значению генеральной
совокупности...**

**...вот в таком
прекрасном
виде!**



**Из нескольких следующих глав мы узнаем,
почему это имеет такое значение.**

**Это знание наконец
дает нам что-то,
в чем мы можем
быть уверены!**



**Мы не знаем
всего...**

**...но это
не означает,
что мы не знаем
ничего!**



Глава 8

ВЕРОЯТНОСТИ

А вот теперь
мы можем начать
нашу охоту!

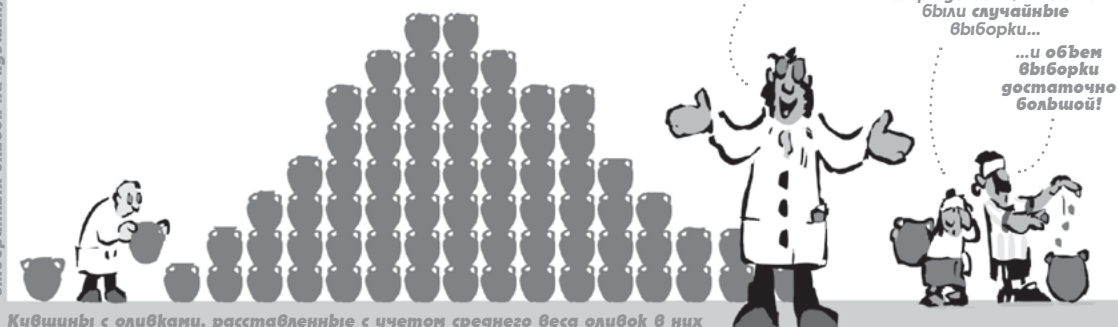


Из предыдущей главы мы узнали, что огромное множество средних значений выборки...

...обычно стремится к нормальному распределению.

Мы можем заявлять об этом с уверенностью...

Количество кувшинов (1500 случайным образом отобранных оливок на кувшин)



Мы собираемся узнать, почему же это так важно...

И что такого необычного в этой форме?



...изучив огромное множество средних значений выборки...

...в сарайчике Безумного Билли, где он хранит свои снасти.

Привет!
Я Билли!

Будьте осторожны...

...он сумасшедший!



Безумного Билли
так называют, потому что
он проводит сумасшедшее
количество времени, создавая
случайные выборки червей...

По 30 червей
в выборке.

Отбирал я их
совершенно случайным
образом...

...из всех червей
в болоте.



**...складывая каждую выборку
в консервную банку...**

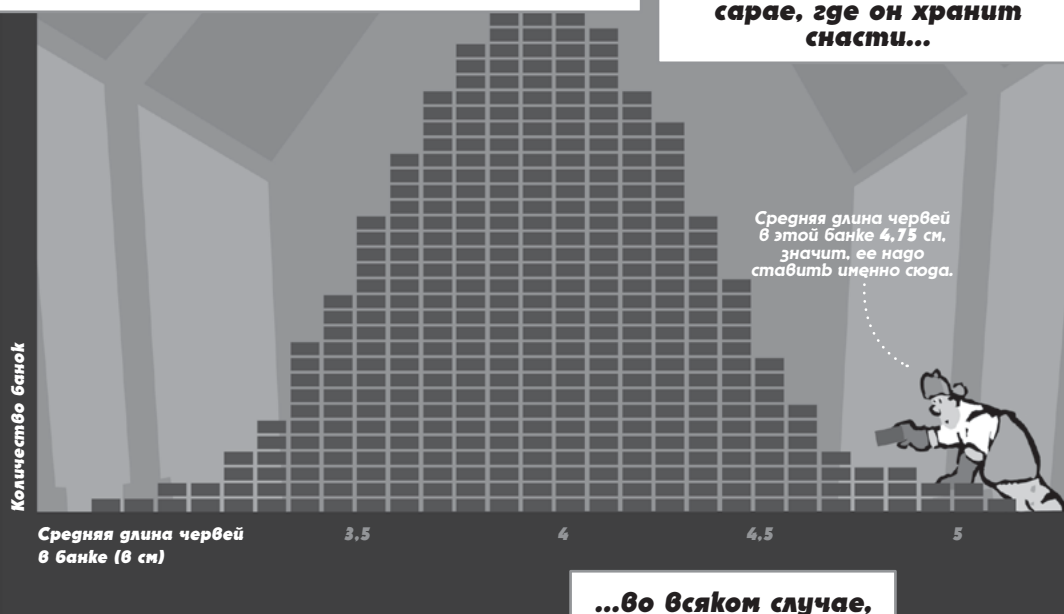
Прежде чем
запечатывать банку,
я замеряю всех червей...

...и вычисляю
среднюю длину червя
в каждой банке.



**...и составляя одну на другую сто тысяч
миллионов таких банок, где каждая
соответствует своему среднему
значению...**

**...в своем безразмерном
сараяе, где он хранит
снасти...**



**...во всяком случае,
так он уверяет.**

Так у тебя что же,
есть настоящее
распределение
выборки?

Ну да, вон за той
дверью.



В этой главе мы выясним, что можно узнать об огромном множестве, которым располагает безумный Билли.

Оно имеет нормальное распределение!

Центрировано по отметке в 4 см.

Стандартное отклонение равняется 0,25 см.

И что?

Более подробно мы остановимся на следующих вопросах:

если у нас будет доступ только к тому, что внутри сарайчика...

...что же мы сможем сказать о генеральной совокупности червей в болоте?



Что же банки с червями, собранными Билли...

...говорят нам об остальных червях, все еще живущих на свободе?

А вот тот же вопрос, но научными терминами.

Если у нас есть распределение выборки, сделанное по средним значениям...

...что мы можем сказать о генеральной совокупности, которую изучаем?



**Первый важный вывод,
который мы бы сделали, если бы
нам удалось хотя бы мельком
заглянуть в сарайчик Билли...**

**...касаясь бы среднего значения
генеральной совокупности.**

Помните, что в
конце концов средние
значения выборки
стремятся к среднему
значению генеральной
совокупности.

Таким образом, среднее
значение признака
совокупности всего
болота оказывается
ровнехонько в середине
этого огромного
множества!

Как раз вот здесь,
возле отметки в 4 см.

Средняя длина червя
в банке (в см)

3,5

4

4,5

5

**Иными словами, если бы нам нужно было
вычислить среднее значение генеральной
совокупности в болоте...**

**...мы могли бы просто
заглянуть в сарайчик —
и нашли бы его там!**

Какова
средняя длина
червей в этом
болоте?

Нет нужды
пачкать одежду,
ковыряясь в этой
грязи.



Но это еще не все...

**Другим важным
открытием, которое
мы могли бы сделать,
заглянув в сарайчик Билли...**



А это уже
и правда
очень важное
открытие!

**...был бы подсчет вероятностей
в отношении генеральной совокупности!**



Что такое
вероятность?

Это просто красивое
слово, означающее
«возможность» или
«шанс».



И вот как это работает:

**если бы мы могли подсчитать
все консервные банки
в огромном множестве,
которым располагает Билли...**

**...и обнаружить, что у 50% из них
среднее значение колеблется в этих
пределах...**

Помните, что
в каждой банке
30 случайно
отобранных
червей.

Все банки
в закрашенной
части графика...

...со средней длиной
3,75 и 4,25 см.



Средняя длина червя
в банке, см

**...это бы означало, что, заведи
мы одну случайную банку
из совокупности...**

**...с 50%-ной вероятностью ее среднее
значение находилось бы в тех же пределах!**

30 случайно
отобранных
червей как раз
на подходе!



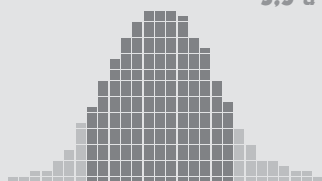
Можно сказать, что с 50%-ной
вероятностью средняя длина
червя будет колебаться
между 3,75 и 4,25 см!



Если бы мы подсчитали
все консервные банки
и обнаружили, что в множестве,
собранном Билли:

95% всех банок...

колеблются между
3,5 и 4,5 см!



Это бы означало, что
в генеральной совокупности:

есть **95%-НАЯ** вероятность,
что средний показатель
в следующей банке, которую
мы заполним наугад собранными
червями из болота...

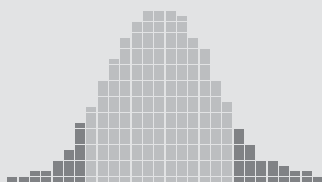
...будет колебаться
между 3,5 и 4,5 см!



А если предположить,
что в этом множестве:

у 5% всех банок...

средний показатель
меньше 3,5 и больше
4,5 см!



...то можно было бы сделать **такие**
выводы о совокупности:

Существует **5%-НАЯ** вероятность,
что среднее значение в банке,
которую мы заполним случайно
отобранными червями из болота...

...будет меньше 3,5
и больше 4,5 см!



Иными словами,
заглянув мельком
в сарай...

...мы можем посчитать,
каков диапазон средних значений
экземпляров, собранных
со всего болота!

Мое множество
все равно
что хрустальный шар
для предсказаний!

С его помощью я могу
сказать, банку с каким
средним значением вы,
возможно, возьмете
следующей!



**Есть несколько вещей,
о которых нужно помнить
при подсчете вероятности*.**

Ом-м-м-м...



* См. стр. 218.

**Во-первых, вероятности актуальны только
в долгосрочной перспективе...**

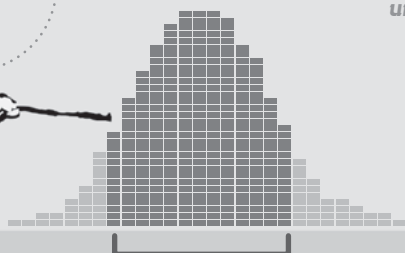
**...поэтому они никогда не скажут ничего достоверно
о коротком периоде.**

Например, если есть
95%-ная вероятность,
что средний показатель
следующей банки, которую
мы наполним случайными
червями из болота...

...будет
равен числу,
колеблющемуся
в этих пределах...

...это не означает,
что и у следующей
банки обязательно
будет среднее
значение из этого же
интервала!

Это означает,
что вероятность
очень высока,
потому что
в долгосрочной
перспективе
у 19 банок из 20
все именно так!



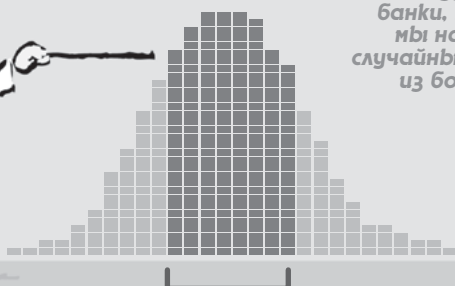
**Во-вторых, у каждой вероятности
есть обратная сторона...**

Например, если есть
50%-ная вероятность,
что средний показатель
следующей банки,
которую мы наполним
случайными червями
из болота...

...будет равен числу,
колеблющемуся
в этих пределах...

...есть ровно
такая же 50%-ная
вероятность, что
средний показатель
следующей
банки, которую
мы наполним
случайными червями
из болота...

...будет
равен числу,
выходящему
за этот
предел!



...потому что вероятности составляют 100%.

Всегда есть
50%-ная вероятность,
что случится какая-нибудь
история...

...при этом есть
и другая 50%-ная
вероятность,
что произойдет
что-то другое.

Если допустить,
что что-то произойдет
с вероятностью
в 95%...

...всегда будет
5%-ная вероятность,
что произойдет
и что-то другое.



И наконец, мы, по определению, можем высчитать, с какой вероятностью произойдут события, только если они происходят случайно...



Вероятность, по определению, означает степень возможности наступления определенного события в долгосрочной перспективе.

...вот почему мы собираем статистические данные только случайным образом.



Если бы я не собирал своих червей случайным образом...

...множество в моем сарае не имело бы никакого смысла.

Говоря общо, мы можем высчитать вероятность других случайных событий, например при подбрасывании монетки...



Вероятность того, что, подбрасывая монетку, вы получите решку...

...составляет 50%...

...потому что в долгосрочной перспективе мы можем предположить, что в 50% случаев будет выпадать решка.



...или броске игральных костей.



Вероятность того, что, бросив кости, вы получите шестерку...

...составляет 1/6...

...потому что в долгосрочной перспективе 1/6 всех вариантов выпадает на выходе шестерку.



Но давайте-ка вернемся к поиску червей случайным образом...



Надеваем на глаза повязку!

...потому что нам предстоит узнать еще кое-что важное о сарайчике Билли!

Оказывается, нам совсем не обязательно пересчитывать все банки в сарайчике Билли...



Сто тысяч миллионов один...
Сто тысяч миллионов два...
Сто тысяч миллионов три...

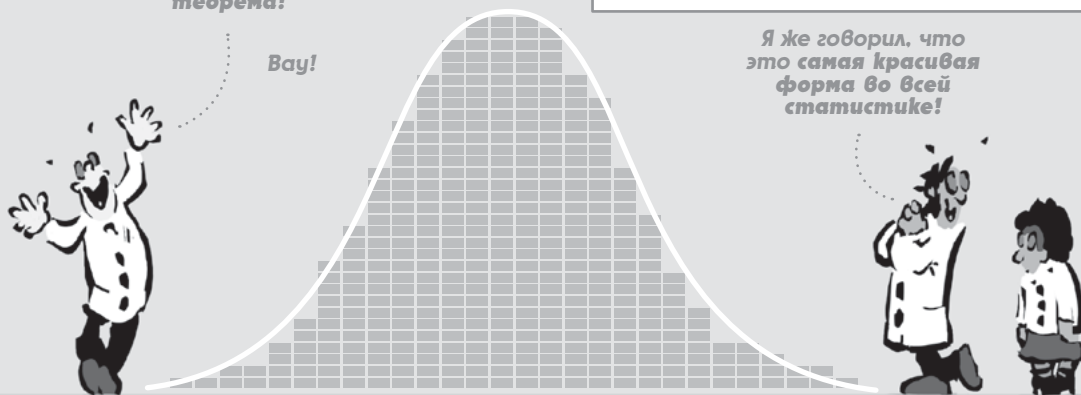
...чтобы высчитать вероятность.



У 95% всех моих банок
средний показатель
колеблется между
3,5 и 4,5 см!

Оказывается, зная, что все множество банок Билли имеет нормальное распределение...

Это и есть
центральная предельная
теорема!

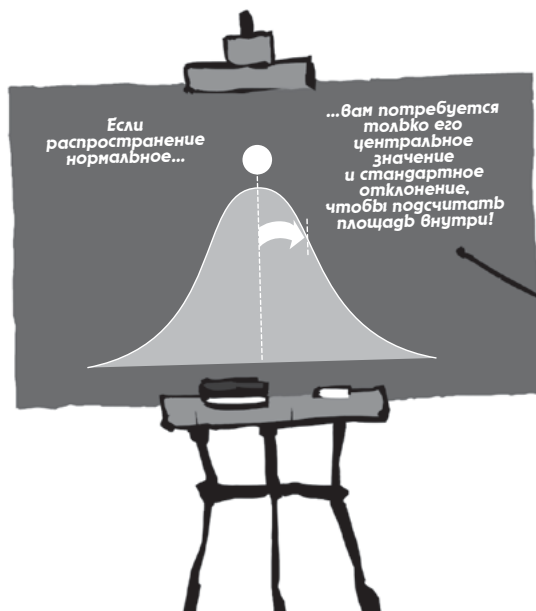


Средняя длина червя
в банке, см

Я же говорил, что
это самая красивая
форма во всей
статистике!

Более того, именно потому что множество распределено нормально...

...нам нужно всего лишь знать его центральное значение и стандартное отклонение, чтобы все высчитать*.



Если
распространение
нормальное...

...вам потребуется
только его
центральное
значение
и стандартное
отклонение,
чтобы подсчитать
площадь внутри!

* Если вы любитель
математики,
см. стр. 219,
там будет больше
объяснений.



А вот настоящие подсчеты, которые предполагает классическая математика, на самом деле очень сложны.

Настолько, что статистики их даже не делают.

Привет, компьютер!



К счастью, есть так называемое правило большого пальца, которое действует в случае любого нормального распределения:

Мы считаем, сколько стандартных отклонений имеем относительно центра.



Мое множествоцентрировано по отметке 4 см, а стандартное отклонение составляет 0,25 см.



68% всех консервных банок...

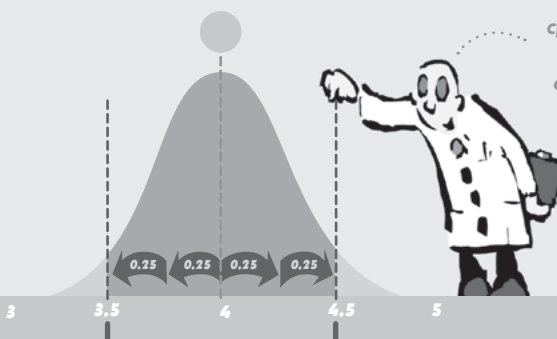
...находятся в пределах 1 стандартного отклонения от центра.



В данном случае среднее значение находится в диапазоне от 3,75 до 4,25 см.

95% всех консервных банок...

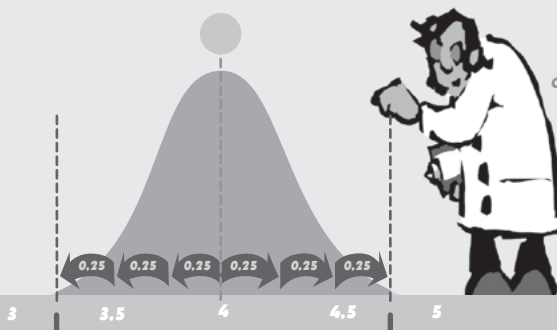
...находятся в 2 стандартных отклонениях от центра.



В этом случае среднее значение находится в диапазоне от 3,5 до 4,5 см.

99,7% всех консервных банок...

...находятся в 3 стандартных отклонениях от центра.



В этом случае среднее значение находится в диапазоне от 3,25 до 4,75 см.

Если создается впечатление, что все эти цифры только сбивают с толку...



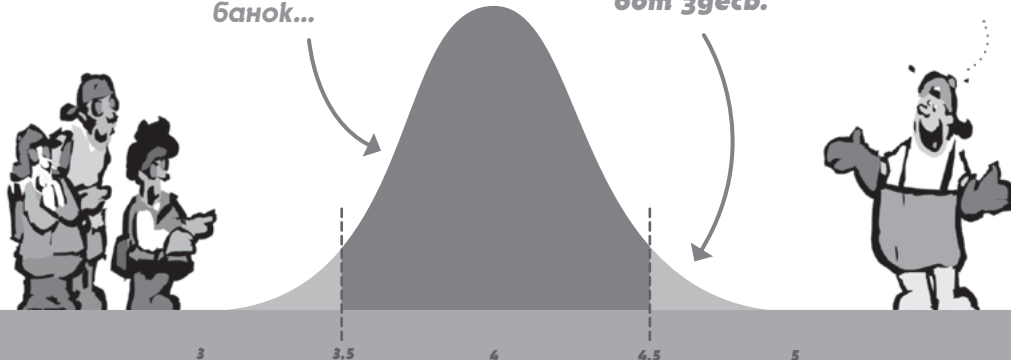
От цифр я цепенею, доктор.

...просто сконцентрируйтесь на затемненных областях.

Очевидно, что внутри этой затемненной области, представляющей собой часть множества Билли, намного больше банок...

...чем вот здесь.

Этот купол больше, чем его хвосты!



Что тут важно помнить, так это то, что затемненные области внутри распределения выборки Билли...

...напрямую соотносятся с нашими шансами собрать средние значения из болота!



Именно поэтому статистики так любят распределение выборки!

Давайте-ка
резюмируем.

Первая замечательная вещь, которую мы узнали о распределении выборки Билли...



Какова же средняя длина
всех червей в твоём
болоте, а, Билли?

...это то, что оно показывает нам среднее
значение генеральной совокупности!

Ответ
вы найдёте
в моём сарайчике!



Вторая замечательная вещь, касающаяся распределения выборки Билли...

...это то, что мы можем использовать его, чтобы
высчитывать вероятность для всей генеральной
совокупности...

Так как мы знаем,
что она
нормальная...

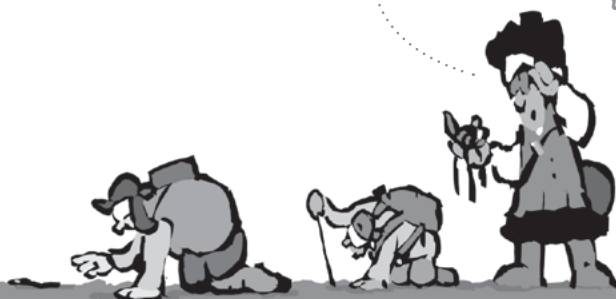
...все, что нам
нужно узнать,
это центральное
значение...

...и стандартное
отклонение!

Если мы пойдём
и сделаем другую
случайную выборку
из 30 червей из того
болота...

...какова вероятность,
что их среднее значение
будет колебаться между
3,75 и 4,25 см?

Позвольте мне
заглянуть
в мой сарайчик,
и я вам скажу!



Ясно, что, если бы мы охотились
за средним значением во всей
генеральной совокупности...

Оно должно
быть где-то
здесь.



...распределение выборки по типу того, что было
в сарайчике Безумного Билли...

Формально
мое распределение
выборки...

...представляет
собой особый вид
вероятностного
распределения!



...было бы для нас невероятно
полезным.

**Все,
довольно!**

**Я хочу посмотреть
на это!**

**Это же
золотая
жила!**

Мое множество средних
значений похоже
на хрустальный шар...

...ты можешь
всмотреться
в него, а увидишь
информацию
о генеральной
совокупности!



К несчастью...

Что?!

Там ничего нет!
Пусто!

...оно не существует.

На самом деле **НЕТ** такого
распределения выборки,
на которое можно было бы
взглянуть!

Как показывает
практика...

Это все
плод моего
воображения.

Я помню каждую
консервную банку,
которую когда-либо
продавал.

...все, что мы можем получить,
это одну банку.

&#@%!



Ну, так что,
какова средняя длина
всех червей в болоте?

Эм-м...



Глава 9

СТАТИСТИЧЕСКИЙ ВЫВОД

Думаю, лучше бы
нам ее
открыть.



**Ясно как день, что у нас
по-прежнему есть
нерешенная проблема...**

*Есть у кого-нибудь
консервный нож?*



**...и она сводится
к следующему:**

**мы пытаемся обнаружить нечто,
чего не можем увидеть.**



**Невозможно,
заглянув в одну
выборку...**

**...увидеть
среднее значение
в генеральной
совокупности.**

*Мы всего лишь
30 жалких червяков.
В то время как
в болоте можно
найти еще сотни
тысяч миллионов.*



**Как будто мы бредем
в тумане на ощупь,
пытаясь найти снежного
человека.**

**Я верю всем
сердцем, что он
где-то там!**

**Но все равно
вы никогда
не сможете его
найти.**

**К счастью, хотя мы и не можем увидать то,
что ищем...**

**Ничего не вижу
за туманом.**

**...мы можем продолжать искать
по подсказки...**

**...которые помогут нам понять, где то,
что мы ищем, может находиться.**

**Если бы ты был
средним значением
совокупности,
где бы ты прятался?**

**Под куполом
того холма!**

Когда мы
пытаемся угадать
местонахождение
среднего значения
генеральной
совокупности...



...мы можем опираться в своем
предположении на что-то, в чем уже уверены...



...и мы с вами даже уже выяснили, что это такое:

В конце концов,
случайные средние
значения выборки,
как правило,
скапливаются вокруг
среднего значения
в генеральной
совокупности...



...и обретают вот
такую красивую
форму!

Это называется
центральная
предельная
теорема!

Вау!



Вот что мы сейчас будем делать:

поскольку
средние значения
выборки обычно
скапливаются...

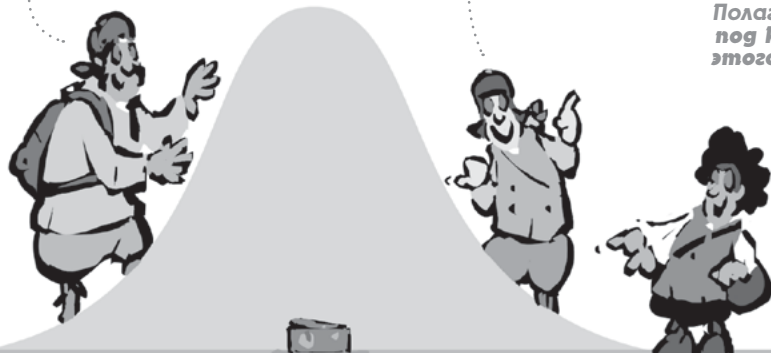
...мы можем нарисовать
вот такую
возвышенность...

...вокруг среднего
значения генеральной
совокупности...

...чтобы угадать,
где находится среднее
значение генеральной
совокупности.

...вроде этого...

Полагаю, оно
под куполом
этого холма!



**Статистики называют этот процесс
статистическим выводом...**

Мы не можем увидеть
его собственными
глазами...

Это похоже
на охоту за снежным
человеком...

...поэтому ищем
те значения,
которые, как нам
кажется, будут
концентрироваться
вокруг него.

...когда выходишь
на след, обнаружив
отпечатки
огромного размера.



**...мы собираемся посвятить всю оставшуюся главу тому,
чтобы в общих чертах обрисовать первый шаг.**

Мы хотим
использовать
одну выборку...

...чтобы представить
себе, что бы мы увидели,
если бы отобрали гораздо
больше экземпляров
для исследования.



Наша основная цель на данном этапе...

...создание иллюстрации.



Я же говорил тебе,
в статистике
главное —
нарисовать
картинку!

На этой картинке видны,
как нам кажется, средние
значения выборки...

И помни,
мы нарисовали
все это...

...потому что
мы охотимся
на среднее значение
в генеральной
совокупности.

**...если бы мы пошли и собрали
сто тысяч миллионов
образцов.**

Эй, вы где там
прячьтесь?



**И в основе всего этого лежит информация,
которую мы получаем из одной случайной выборки.**

Какой у тебя
объем выборки?

...А размах вариаций?

...А расположение?



Как будто мы делаем свое
самое смелое предположение
о том, как выглядит
воображаемое множество
Билли...



...имея в руках только
одну консервную
банку.



Наши действия очень похожи
на магию...



Я кладу одну
банку червей
в шляпу.

В рукавах
ничего нет.

Фокус-покус! Перепokus!

И из одной
получается
так много!

Сто тысяч
миллионов банок!



...Хотя на самом деле все
очень просто.

Чтобы сделать нужный рисунок, мы используем центральную предельную теорему в качестве плана:

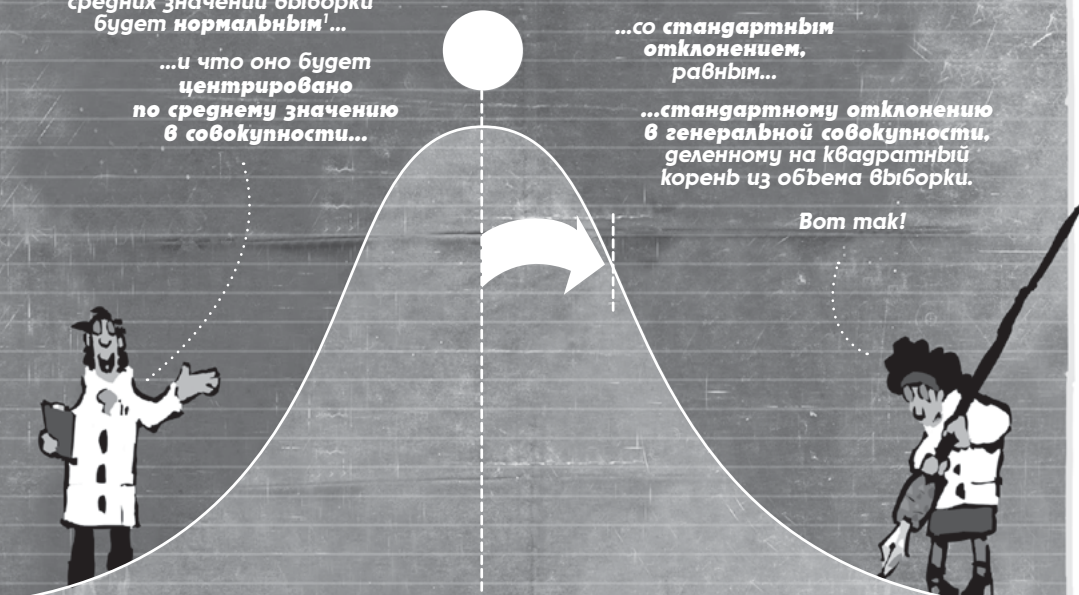
мы можем предположить, что огромное множество средних значений выборки будет нормальным!...

...и что оно будет центрировано по среднему значению в совокупности...

...со стандартным отклонением, равным...

...стандартному отклонению в генеральной совокупности, деленному на квадратный корень из объема выборки.

Вот так!



1. Помните, что тут возникают некоторые ограничения, см. стр. 102.

Поскольку мы не знаем настоящих значений в генеральной совокупности...

И никогда не узнаешь!



...мы просто заменяем их теми, которые получили из нашей выборки.

Давай сделаем вид, что среднее значение в твоей банке такое же, как среднее значение во всем болоте!

И размах вариаций в твоей банке такой же, как в болоте!

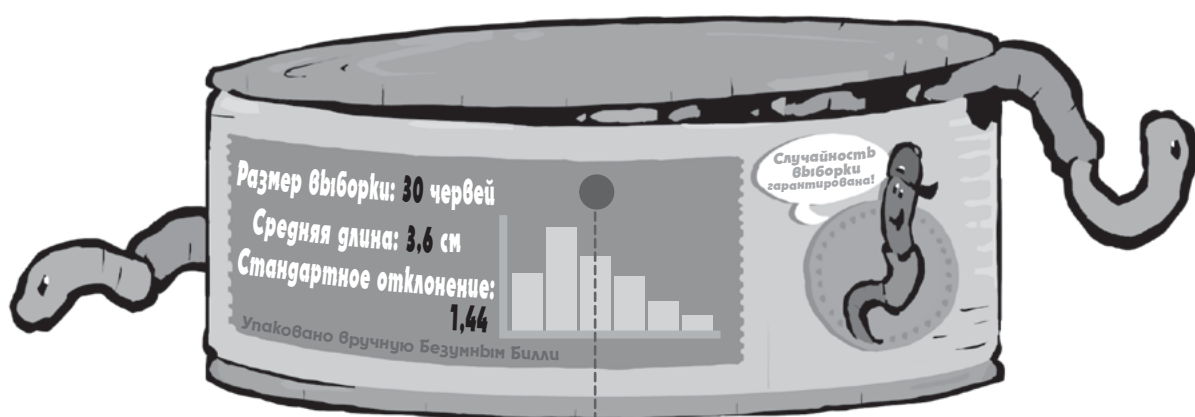
Разве это не мухлеж?

Нет, это всего лишь наше самое смелое предположение!

Мы предпочитаем называть это аппроксимацией.



Так, например, когда мы используем
выборочные значения из одной банки...



...мы рисуем картинку, которая выглядит примерно так:

Наше предполагаемое
огромное множество
средних значений
нормально распределено...

...и центрировано
по среднему значению
в одной консервной банке...

3,6

...со стандартным
отклонением,
равным...

...нашему стандартному
отклонению, поделенному
на квадратный корень объема
выборки!

Вот так!

$$CO = \frac{1,44}{\sqrt{30}}$$

Мы называем эту картинку предполагаемым выборочным
распределением*.

Это предположение...

...о том, как средние
значения выборки
могли бы быть
распределены...

...если бы
мы насобирали
их целую тонну.

* См. стр. 219, чтобы
узнать, как описывать
подобные случаи, используя
математические символы.

Теперь, используя
одну выборку...

...чтобы создать предварительное
распределение выборки...

Имея на руках только
30 случайным
образом отобранных
червей, мы можем
предположить...

3.6

...как бы выглядело
огромное множество,
собранное безумным
Билли, если бы оно
существовало!

CO=0.26

Неплохо, да?

...мы можем подвести итоги нашей охоты на среднее
значение генеральной совокупности.

Ну так что, оно
под куполом
холма?

Или все же
нет?

Как ни странно,
мы можем быть
уверены, что так
оно и есть.

А в следующей
главе узнаем,
насколько
мы можем быть
в этом уверены.



Глава 10

ДОСТОВЕРНОСТЬ

Учитель, я полон
сомнений...

Что мне
делать?

Тебе нужно
поучиться
статистике!



**Помните, что в результате
нам нужно узнать что-нибудь
о среднем значении
в генеральной совокупности.**

Мне вообще наплевать
на следы гигантского
размера...

...я хочу узнать
что-нибудь о самом
снежном человеке!



**К сожалению, несмотря
на все те магические трюки,
которым мы только
что научились...**



У меня нет
ничего
в рукавах...

...а в руке только
консервная банка
с червями.

**...мы никогда не сможем
добиться этого.**

Нет никакой возможности
заглянуть в консервную
банку и увидеть среднее
значение в генеральной
совокупности...

...равно как и нет никакой
возможности заглянуть
в предварительное
распределение выборки
и увидеть то, что мы ищем.



Никогда!

Эх!



**Вот поэтому мы и учимся делать
предположения.**

Я никогда
не смогу
найти то,
что ищу!

Не отчаивайся!

Ты всегда можешь
высказать догадку
относительно
месторасположения
этого среднего
значения!



Мы пока поговорили только о том, как выглядит первый шаг в процессе выстраивания предположений...

...но нам еще предстоит разобраться со вторым.

Не торопясь, со всей любовью, изобразите предварительное распределение выборки...

...обратив особо пристальное внимание на самые важные детали.

Порежем это на кусочки!



Итак, в этой главе мы научимся детализировать наш рисунок...

...состригая аккуратненько по краям...

...и используя то, что осталось, чтобы вычислить степень достоверности.

Так намного лучше!

Теперь я могу быть уверен в том, что именно я вижу.



Раз уж мы уже знаем,
как нарисовать предварительное
распределение выборки...

Как же красива
эта форма!

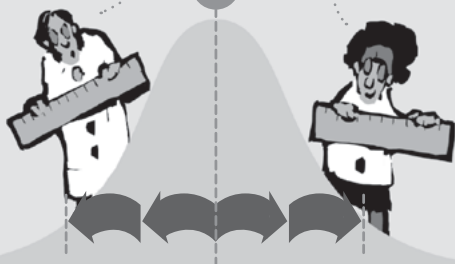


...мы легко научимся высчитывать и степень достоверности.

Нужно просто хорошенько
взглянуть в то,
что мы только что нарисовали...

На этот раз нам нужно отмерить
2 стандартных отклонения
от центрального значения...

...по обеим
сторонам.



...и отрезать все «хвостики»!

Вжук-вжук!

Сидите смирно,
это совсем
не больно.

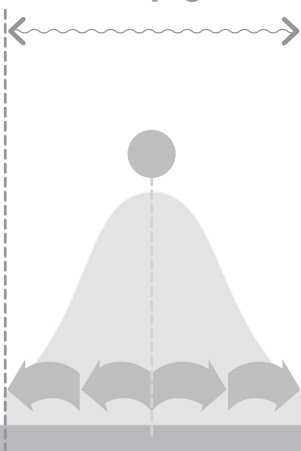


А потом мы делаем утверждение,
например такое:

мы уверены
на 95%...

...что среднее
значение генеральной
совокупности
находится где-то
в этих пределах!

Мы еще
вернемся
к этим
«хвостикам»
в главе 11.



Ура!

**Только
и всего!**

**Отмеряйте
и отрежьте!**

Тут даже
ребенок
справится!

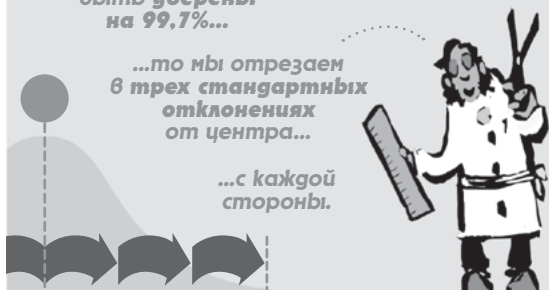


**Если мы хотим большей
достоверности, нам нужно
просто отрезать чуть дальше.**

Если мы хотим
быть уверенны
на 99,7%...

...то мы отрезаем
в трех стандартных
отклонениях
от центра...

...с каждой
стороны.

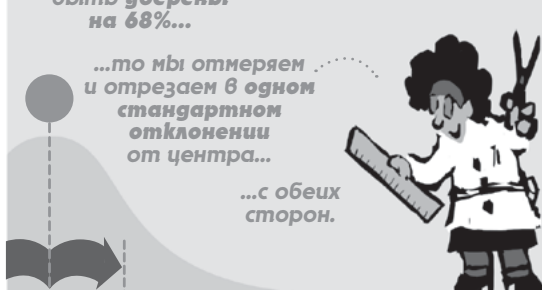


**А если нам нужно меньше
достоверности, то нужно
отрезать чуть ближе.**

Если мы хотим
быть уверенны
на 68%...

...то мы отмеряем
и отрезаем в одном
стандартном
отклонении
от центра...

...с обеих
сторон.



В основе всех этих подсчетов лежит то,
что мы уже изучили на стр. 115!

**Но где бы мы ни отрезали, мы всегда декларируем
нашу степень уверенности с помощью двухчастного
утверждения...**

**Мы уверены
на 95%...**

...что среднее
значение
в совокупности
находится где-то
в этом пределе!

**...в котором объединены
и степень достоверности...**

...и доверительный интервал*.

Мы твердо
уверены
в своей
правоте!



Мы, конечно,
никогда этого
не докажем...

...но мы бьемся
об заклад, что
это где-то
здесь!



* Доверительный интервал представляет собой интервальную оценку. См. стр. 220, если хотите узнать больше.

Например, если мы возьмем предварительное выборочное распределение, сделанное с помощью нашей консервной банки...



Мы проделали это на стр. 129.

3.6

0.26

0.26

0.26

0.26

3.08

3.6

4.12



...и отрежем «хвосты»...



...на расстоянии в 2 стандартных отклонения от центрального значения...

3.08

4.12



...мы сможем сказать следующее:

Мы уверены на 95%...

...что среднее значение в совокупности находится между 3,08 и 4,12 см!



Но что конкретно это означает?

Мы создали всего лишь одно предварительное выборочное распределение с помощью одной случайной выборки...

Мы построили это с помощью одной банки...



3.6

...наполненной 30 червями.

0.26

...и использовали его, чтобы подсчитать один доверительный интервал.

Мы уверены на 95%...

...что среднее значение в совокупности колеблется между 3.08 и 4.12 см.



Но если бы мы взяли другую случайную выборку и использовали ее, чтобы создать другое предположительное распределение выборки...

Мы построили это с помощью другой банки, наполненной 30 червями.



4.1

У нее совершенно другой центральный показатель...

...и размах вариаций.

0.23

Мы уверены на 95%...

...что среднее значение в совокупности колеблется между 3.64 и 4.56 см!



И если бы мы продолжали собирать новые и новые случайные выборки и выстраивать новые и новые предполагаемые выборочные распределения...

Собери и посчитай.



Собери и посчитай.



Собери и посчитай.



Собери и посчитай.



Собери и посчитай.

Собери и посчитай.



Собери и посчитай.



Собери и посчитай.



Собери и посчитай.



...мы бы продолжали получать разные интервалы.

Это важно, потому
что **единственный вывод,**
который мы можем сделать
из этого...



Мы на 95%
уверены...

...что среднее
значение совокупности
находится где-то
в этом пределе!

...это что, если бы мы установили **таким образом сто тысяч**
миллионов разных пределов...



Собери в произвольном
порядке.

Нарисуй картинку.

Отрежь
«хвостики» в 2 со
от центра.

...среднее значение генеральной совокупности
содержалось бы примерно в 19 банках из 20...

В 19 банках из 20.
Получается 95%!



А примерно в 1 банке
из 20 было бы по-другому.

В одной банке из 20
все совершенно
неправильно!

Будем надеяться,
что мы не взяли
именно ту банку.



Другими словами,
когда мы говорим
так:

**мы уверены
на 95%...**

...что среднее
значение генеральной
совокупности
находится где-то
в этом пределе!

**...это означает, что есть
5%-ная вероятность,
что мы заблуждаемся
на сей счет.**

В таком случае среднее
значение генеральной
совокупности на самом
деле где-то в другом
месте...

...и тогда
получается,
что все наши
усилия тщетны!



Печальная правда заключается в том, что любая выборка, которую мы отобрали случайным образом из генеральной совокупности...



...может оказаться обманчивой и ввести в заблуждение.

Мы можем совершенно случайно насобирать...

...30 очень коротких червяков.



А если одна наша выборка дает настолько неоднозначные результаты...

...то и предварительное выборочное распределение, которое мы рисуем на ее основе, тоже окажется неверным.

Если наше распределение будет основано на информации о 30 очень коротких червяках...

...оно сильно сместится влево.



А что если среднее значение и правда находится где-то здесь?



Это серьезная проблема...

...но мы можем избежать подобных трудностей, если всегда будем помнить о более масштабной картинке.

Эта консервная банка могла оказаться пустышкой, введшей нас в заблуждение!



Может, да, но, вероятнее всего, нет.



Подумайте о долгосрочной перспективе!



Даже если одна выборка оказалась обманчивой...

Мы, конечно, можем собирать 30 очень коротких червей...

...совершенно случайно!

...в долгосрочной перспективе станет ясно, что, скорее всего, это не так...

...потому что большинство средних значений в случайных выборках имеют тенденцию группироваться вокруг среднего значения генеральной совокупности!

Кажется, знакомо?

Да ведь это же центральная предельная теорема!

Вау!

Иными словами, средний показатель одной банки может случайно оказаться здесь...

Какие короткие червячки. Это странно!

...или здесь...

Какие длинные червячки. Это странно!

...а такое маловероятно...

Довольно обычные червячки.

В конце концов оказывается, что у большинства консервных банок среднее значение находится под куполом холма.

...и мы можем быть в этом уверены.

Подводя итог, скажем, что понимание статистической достоверности...



Мы уверены на 95%...



...что среднее значение в совокупности находится где-то в этом пределе!



...предполагает, что мы должны держать в уме как продолжительный период, так и короткий промежуток одновременно.

В долгосрочной перспективе наша приблизительная оценка и отсечение «хвостиков» дают прекрасные результаты. И точка.



Если вы возьмете случайную выборку хорошего размера и с ее помощью изобразите предварительное выборочное распределение...

...затем отмерите 2 σ от центра и отрежете «хвостики»...

...в 95% случаев у вас получится предел, в котором будет находиться настоящее среднее значение в совокупности!



Это было доказано математически!

А также опытным путем!



...Но вот в краткосрочной перспективе всегда есть вероятность, что мы схватили не ту банку!

Мы уверены на 95%, что среднее значение в совокупности находится где-то в этом пределе...

...но так ли это на самом деле?

Может, да, а может, нет.

Мы никогда не будем знать это наверняка!



Глава 11

ОНИ НАС НЕНАВИДЯТ



Они хотят
убить нас!

Насколько вы
в этом уверены?



**Когда мы используем
только одну выборку...**

Если взять
за основу
50 случайно
отобранных
русалок...



**...чтобы подсчитать степень
статистической достоверности
всей генеральной совокупности...**

...я могу
с 95%-ной
уверенностью
сказать...

...что рост всех
русалок в этой
лагуне...

в среднем
варьируется
от 7 до 10 см!



...мы делаем что-то удивительное!

Мы делаем предположение...

Кто бы мог подумать,
что русалки такие
крошечные?

Одна случайная
выборка может
сильно запутать...

**...заслуживающее
доверия предположение...**

...но в долгосрочной
перспективе
это утверждение
покажется весьма
сомнительным.




**...о чем-то, что мы не можем увидеть,
а способны только представить себе.**

Йу-ху-ху!

Эй, есть ли там,
внизу, еще русалки?

Он никогда
не узнает этого
наверняка.



**Все начинается
с трех цифр.**

Достаточно
большой объем
выборки...

...среднее значение
выборки...

...и стандартное
отклонение
выборки.

Но помните,
что все получится только
в том случае, если все
ваши русалки будут
отобраны произвольным
образом.

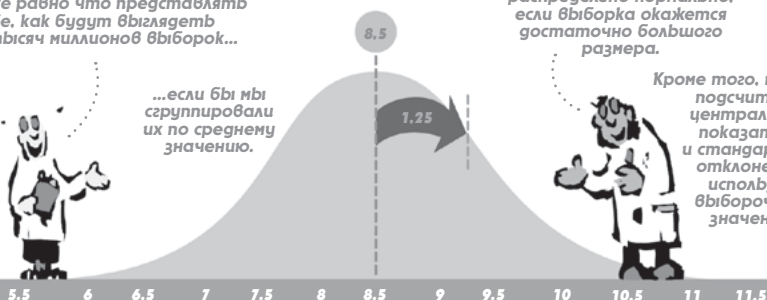
Имея на руках лишь три эти цифры, можно наметить предварительное распределение выборки...

Это все равно что представлять себе, как будут выглядеть сто тысяч миллионов выборок...

...если бы мы сгруппировали их по среднему значению.

Мы знаем, что наше множество будет распределено нормально, если выборка окажется достаточно большого размера.

Кроме того, мы можем подсчитать центральный показатель и стандартное отклонение, используя выборочные значения.



...и отрезать «хвостики»...

Мы считаем от центра, учитывая стандартные отклонения...



...чтобы получить единственное заслуживающее доверия утверждение...

...чтобы выйти за пределы этого массива, где значение вероятности мы знаем наверняка.



В долгосрочной перспективе это отлично работает!

...в котором будут и степень уверенности...

...и доверительный интервал.

Я на 95% уверен...

...что средний рост в совокупности всех русалок этой лагуны — где-то от 7 до 10 см!



Как мы уже знаем, из-за того что этот метод требует изрядного количества подсчетов...

Ну во-о-от!



...он хорош только для тех характеристик, которые можно измерить.

А сколько русалки весят?

А какой они глины?

Сколько у них зубов?

Тебе нужны числовые данные.

Об этом мы рассказывали в главе 4.

Поэтому может показаться, что он неприменим к характеристикам, которые не поддаются явному численному измерению...

Они счастливы?

Все ли они поют красиво?

Если я ткну в них палочкой, насколько им будет больно?

...однако это не всегда так.

Правда заключается в том, что мы можем высчитать степень достоверности относительно любой характеристики...

А русалки по жизни оптимистки?



Интересно, они сообразительны?



А они любят есть суши?



...если найдем способ измерить ее...

Вот тебе тест.



...и сможем отметить показатели на числовой оси.

Если ваш балл за тест находится где-то здесь, вы... гурочки!

Если вы набрали столько баллов, вы гении!



В этой главе мы будем заниматься как раз этим...

...чтобы исследовать вопрос, касающийся ненависти.

Как сильно я тебя ненавижу?

Погоди, я подсчитаю все варианты...



Всем известно, что негодники,
живущие на планете Бип...

...ненавидят хороших людей,
живущих на соседней планете Пип.

Фу-у! Фу-у!

Вы. @*\$&
пипиане!

Они нас
ненавидят!

Да эта
ненависть
измеряется
триллионами
поколений!

И вот вопрос,
который нас мучает:

...правда ли это?

Так как мы не можем
опросить лично
все 785 000 000 000 бипиан,
живущих на планете, о том,
что они чувствуют...

...нам ничего не остается, кроме как
основываться в своих суждениях на случайной
выборке.

Бипиан
слишком
много.

Иными словами,
нам придется
воспользоваться
статистическими
приемчиками.

Помните, никто
не использует
статистические
техники,
если только
в этом нет
нужды!

Но прежде чем мы отправимся
делать случайную выборку...

...нам нужно придумать, каким
образом нам перевести чувства
каждого бипианина по отношению
к пипианам...

...на язык цифр.



Я бы оторвал
этим пипианам
голову!

Нам нужен
цифровой
эквивалент
для этого.



Никому не говори...

...но я думаю, они
милые ребята!



Давай превратим
все эти голые
эмоции в цифры!



В этом случае мы можем придумать
свою систему исчисления...

...которая будет варьироваться от чистой ненависти...

...до истинной
любви.

Мне хочется убить
каждого пипианина,
которого я вижу!

Лично мне
они не очень
нравятся.

Честно говоря,
мне как-то
все равно.

А по-моему,
они
классные.

А моя мечта —
жениться
на какой-нибудь
симпатичной
пипианочке.



-10



-5



0



5



10

Давай переведем слова каждого бипианина, с которым поговорим...

...в числовой эквивалент по шкале от -10 до 10.



Я готов переломать
им кости и буду
только рад этому!

Понятно:
ваш балл
-10.





Как бы вы оценили
свою ненависть
к пипианам по шкале
от -10 до 10?

Как бы вы оценили
свою ненависть
к пипианам по шкале
от -10 до 10?

Как бы вы оценили
свою ненависть
к пипианам по шкале
от -10 до 10?

**Затем мы аккуратно
соберем все данные,
которые...**

**...получили произвольным
образом...**

Вообще, между
выборкой
бипиан, которых
мы опрашиваем...

...или любой
другой
выборкой...

...может
и не быть никакой
систематической
разницы...

...поэтому давайте-ка
поищем по всей планете
и отберем бипиан
совершенно случайным
образом...

...немного
отсюда...

...нескольких
отсюда...

...чуть-чуть
отсюда...

...и здесь
поищем...

...немного
отсюда...

...немного
отсюда...

...и т. д.

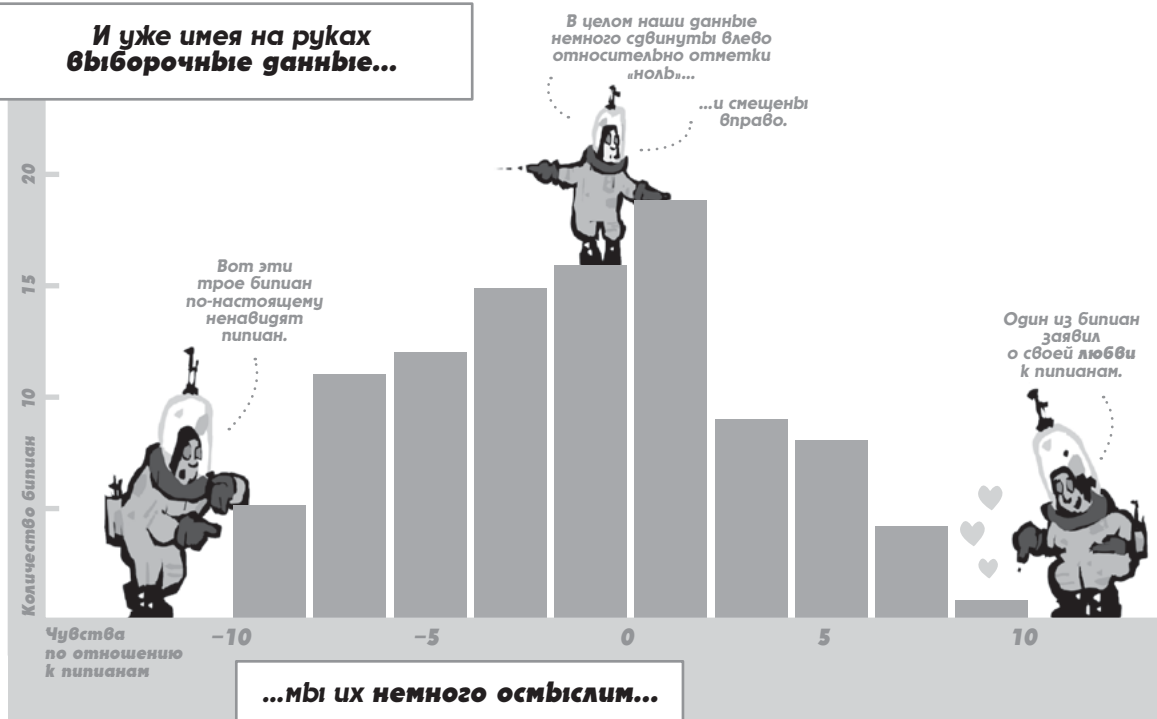
...и т. д.

**...памятуя о том, что собрать
нам нужно достаточное
количество...**

**...чтобы быть уверенными,
что все подсчеты верны.**

...и не остановимся,
пока не отберем
100 бипиан.

И уже имея на руках выборочные данные...



...и выделим три параметра, которые нам понадобятся, чтобы сделать
статистический вывод.

**Объем
выборки
равен 100.**

Выборка
достаточно
размера, чтобы
можно было
высчитать
уровень
достоверности.



**Выборочное
среднее
значение
равно -1.**

Негативно,
но не сильно.



**Стандартное
отклонение
выборки
равно 4.**

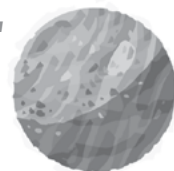
Похоже, вариантов
слишком много,
учитывая тот
факт, что
шкала имеет
всего 20 делений
в ширину.



Мга-а...



Какие же выводы мы
можем сделать насчет
185 000 000 000 с лишним
бипиан в генеральной
совокупности?



**Итак, еще раз, для начала
нас интересуют три показателя.**

**Объем выборки,
которую мы сочли
достаточно большой,
равен 100...**

Если мы не можем
сделать выборку
достаточно объема...

...нам придется
воспользоваться
специальными
инструментами,
о которых мы узнаем
в главе 14.

**...выборочное
среднее значение
равно -1...**

**...а стандартное
отклонение
выборки равно 4.**



**С помощью этих трех показателей
мы намечаем предполагаемое
распределение выборки...**

Оно нормальное...

...и центрировано
по среднему
значению выборки.

Ура-а-а!



-1

Стандартное
отклонение
равно...

...стандартному
отклонению нашей
выборки...

...поделенному на
квадратный корень
объема выборки.



$$\frac{4}{\sqrt{100}}$$

-2.0

-1.0

0.0

...и отрезаем «хвосты»...

Мы отсчитаем
от центра
2 стандартных
отклонения...

...и получим
95% площади
внутри нашего
пика.



А потом посмотрим
на весь диапазон
значений.

0.4

0.4

0.4

0.4

**...чтобы получить
единственное заслуживающее
доверия утверждение...**

-1.8

-1

-0.2



...хотя
в краткосрочной
перспективе мы можем
заблуждаться.



**...в котором будут
учтены и уровень
доверия...**

Мы уверены
на 95%...

...и доверительный интервал.

...что среднее
значение
в генеральной
совокупности всех
биопан на планете
находится где-то
в диапазоне между
-1.8 и -0.2!



Итак, что же?

Руки чешутся
выдавить им глаза
пальцами...

...отрезать бы
им их хвосты
тупым ножом...

...ну-ка, выплесните
всю свою злобу...

...и подожгите
пары своей
ненависти!

Ненавидят ли
бипиане пипиан?

Вигел?

Просто какая-то
кучка озлобленных
извергов!

Раз уж мы не можем опросить
всю генеральную совокупность,
мы никогда не получим точный
ответ на свой вопрос.

Но вот наши выборочные
данные...

...весьма однозначно говорят о том,
что не ненавидят!

Мы уверены
на 95%...

...что настоящее
среднее значение
в генеральной
совокупности
находится
не где-то там,
в зоне ненависти...

...а где-то
здесь!



-1.8

-0.2

А тут мы,
в лучшем случае,
наблюдаем мягкое
непрятие...

...но
...граница не ненависть!
с безразличием...



-10

-5

0

5

10

Помните о том, что максимум, который мы можем предложить с помощью статистики, это портрет с некоторыми нюансами...



Держите в уме и долгосрочную перспективу...

...и краткосрочную!

...потому что любое заключение, сделанное на основании данных выборки...

Это касается всех заключений, сделанных с помощью статистики.

Где угодно, когда угодно, неважно!



...может быть в корне неверным.

Когда мы на 95% уверены, что это здесь...

...мы одновременно на 5% уверены, что это не так!



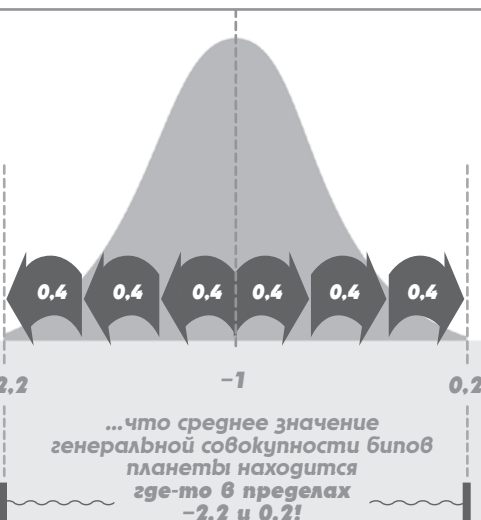
И даже если мы расширим наш уровень достоверности...

...чтобы охватить больший интервал...

Если мы отсчитаем от центра 3 стандартных отклонения, мы сможем сказать:



Мы уверены на 99,7%...



...все равно есть вероятность, что мы ошибаемся.

Наша выборка может оказаться обманчивой...

...но это маловероятно.

Тут, правда, нужно сказать еще об одном...



**Мы только что высчитали
доверительные интервалы,
равные 95% и 97,7%...**

**...основываясь всего на одной
произвольной выборке в 100 бипиан.**

**Мы очень даже
уверены...**

**...в том,
что вы нам
не очень-то
по душе!**

**Но есть еще кое-что,
что мы могли бы сделать,
чтобы быть еще более
уверенными.**

**Никогда
нельзя быть
достаточно
уверенным.**

**Больше
всегда
лучше!**

Если бы мы изначально опросили больше случайно отобранных бипиан...

**Как бы
вы оценили
свои чувства
к бипианам
по шкале
от -10 до 10?**

**Давайте
опросим
больше
100 бипиан!**

**И не будем
останавливаться,
пока не наберем
225!**

**...наше предварительное выборочное распределение могло
оказаться более узким...**

**Посмотри,
что происходит,
когда мы увеличиваем
объем выборки
со 100 до 225
участников.**

**Все множество
становится
более узким!**

**Мы предсказывали,
что так оно
и будет,
на стр. 98.**

-1

$\frac{4}{\sqrt{100}}$

0,4

-1

$\frac{4}{\sqrt{225}}$

0,26

**...а это, в свою очередь, сузило бы
наши доверительные интервалы, а значит,
и сделало их более точными!**

Чтобы посмотреть, как это работает, вспомните, что мы начали с этих трех показателей.

Объем выборки равен 225.

Бипан стало больше в два с лишним раза.

Выборочное среднее значение равно -1.

Стандартное выборочное отклонение равно 4.

Маловероятно, что объем нашей выборки, а также стандартное отклонение останутся такими же и в большей выборке...

...но давайте возьмем то же число и посмотрим, что будет, если мы поменяем только объем выборки.

Так же, как и раньше, с помощью этих трех цифр мы можем наметить предполагаемое распределение выборки...

Это нормально...

...и центрировано по нашему выборочному среднему значению...

...но в целом на этот раз все уже, потому что объем нашей выборки был больше.

...и отрезать «хвостики»...

Давайте отсчитаем 3 СО от центра и получим доверительный интервал в 99,7%.

-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5

И на этот раз, какой бы конкретно доверительный уровень мы ни выбрали...

...у него будет намного более точный интервал.

На этот раз мы уверены на 99,7%...

...что среднее значение в совокупности находится где-то между -1,78 и -0,22.

Это почти то же самое, что и 95%-ный доверительный интервал, который мы получили при опросе 100 бипан!

Вот, собственно, почему больший объем выборки всегда лучше!

Если вы можете увеличить объем выборки...

...сделайте это.

Это только усилит вашу уверенность!

В этой главе мы создали числовую шкалу, благодаря которой можно определить, как одна группа людей относится к другой.

Мой ответ был
оценен в 10 баллов
по предложенной шкале!

Ты выйдешь
за меня
замуж?

**Эту хитрость можно использовать,
чтобы получить ответы на самые
разнообразные вопросы...**

По шкале
от 1 до 10...

...скажи
мне, как
сильно
болит.

По шкале
от 0 до -100...

...оцени,
насколько
ворчлива
ты бываешь
по утрам?

По шкале
от 1 до 100 000 000 000...

...оцени, как сильно
ты меня любишь?

...потому что, если мы соберем достаточно числовых выборочных данных...

**...мы сможем высчитать
степень достоверности
в отношении любой
совокупности...**

Ты считаешь
меня
симпатичным?

...находящейся вне пределов досягаемости.

Глава 12

ПРОВЕРКА ГИПОТЕЗ

*Моя гипотеза в том,
что **ты** жульничаешь!*

*Сначала предъяви
доказательства!*



В нескольких последних главах мы учились высчитывать достоверность...



...намечая предполагаемое распределение выборки...



...и вырезая большой массив в самом центре этого распределения.

Мы уверены на 95%...



...что среднее значение генеральной совокупности находится где-то в этом множестве!



В этой главе мы познакомимся
с новой техникой.

Мы возьмем нашу
предварительную
оценку...

...и посмотрим, что она может сказать,
если мы сдвинем ее на новое центральное
место.



Все это часть процесса,
который носит
название «**проверка
гипотезы**»...*

Это такой
тест...

...который показывает,
правда ли мы думаем,
что среднее значение
генеральной совокупности
может быть...

...ПРЯМО
ЗДЕСЬ!



...и является еще одной важной стратегией
для формирования статистических выводов.

* См. стр. 221, если вас интересует более
подробное описание того,
о чем мы говорим в этой главе.

**«Проверка гипотезы»
звучит забавно...**

Они называют
меня *Реджинальдом*
Предположителем
Джонсом Третьим...

...и да, мои носки
стоят дороже
твоей куртки.



**...но на деле это просто еще
один способ поохотиться
на неуловимое среднее значение
генеральной совокупности.**

**Как мы уже знаем, нам никогда
не удастся увидеть среднее
значение в генеральной
совокупности своими глазами...**



Те-с!



**Надень
повязку!**



**...но оказывается, мы можем немного продвинуться
в нашем квесте, попытавшись определить
его точное местонахождение.**

Эй, народ, а что,
если оно было
именно тут!

Прямо у нас
под носом!



Что мы в итоге будем делать
с проверкой гипотез?

Проверить наши догадки...

Думаю,
это оно!

А что, если это
и правда оно?

...сравнивая их со средним
значением в выборке, которое
мы на самом деле нашли.

Если это
действительно так,
что ты будешь с ним
делать, а?



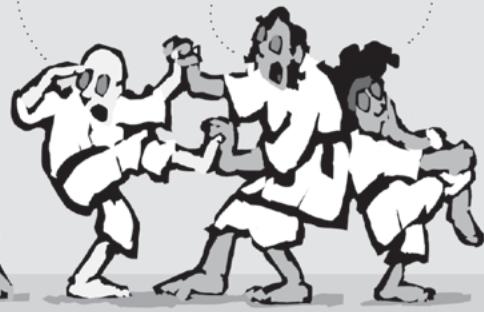
Вообще, вся эта проверка подразумевает, что нужно будет
еще не раз **покрутить** и **повертеть** найденные значения...

Оцени
ситуацию! **Отмеряй!**
Толкай!

Переверни
и посмотри
сюда!

Подсчитай!

**Прими
решение!**



...поэтому спешить не будем.



Тихо-тихо.



Процесс начинается с того, что мы намечаем одно предполагаемое распределение с помощью одной выборки...

Это своего рода отображение того, как бы выглядели другие средние значения выборки...

...если бы они были собраны вместе и центрированы над средним значением в нашей выборке.



...и передвигаем его в другое место, про которое как раз и хотим все узнать.

Обычно мы держим в уме какое-то конкретное место...

...но об этом поговорим в следующей главе.



Благодаря этому мы можем увидеть, как выглядели бы другие выборочные средние значения...

...если бы они были собраны воедино и центрированы в этом месте.



**Затем мы смотрим
на среднее значение,
которое нашли
в нашей выборке,
и решаем:**



**если бы среднее значение
генеральной совокупности
действительно находилось
где-то здесь...**

**...какова вероятность того,
что мы бы в произвольном
порядке нашли такую
выборку, как наша?**

**Чтобы узнать ответ, мы можем
воспользоваться теми знаниями,
которые получили о статистической
достоверности!**



**Если наше предположение верно
и настоящее среднее значение генеральной
совокупности находится здесь...**

**...тогда в долгосрочной
перспективе мы можем
предполагать,
что в любой выборке...**

**...среднее значение будет
находиться под куполом
холма.**

В этом
и заключается
наше великое
открытие!

Вау!



**Но если среднее значение
в нашей выборке будет
отличаться в одну сторону...**

Много
коротких.



...или в другую...

Много
глинных.



**...мы сможем резонно предположить,
что наши гипотезы оказались ложными.**

Полагаю, что это
именно то, что
мы ищем!



Да? Тогда почему
исследуемое нами
среднее значение
находится здесь, а?



**Но давайте все же
поконкретнее об этом.**

В долгосрочной перспективе мы предполагаем, что 95% всех средних значений в выборке будут находиться в пределах двух стандартных отклонений от среднего значения генеральной совокупности...

...поэтому вероятность того, что мы наугад возьмем среднее значение в выборке где-то тут...

...или тут...

...составляет примерно 5%.

Ты можешь наугад обнаружить среднее значение в выборке где-нибудь здесь...

...но это очень маловероятно.



**На практике мы сравниваем
наши выборку и предположение,
подсчитывая то, что
называется вероятностью*
(или Р-значением)...**

Если это и есть среднее значение генеральной совокупности...

...то вероятность того, что мы сделаем выборку где-то здесь, составляет 4%.



**...и если оно меньше 5%,
мы подозреваем, что
наше предположение,
возможно, неверно.**

Извини, конечно, но, как по мне, это слишком маловероятно.



* См. стр. 221-222, где вы найдете словарное определение.

**Мы всегда заканчиваем
проверку гипотез,
принимая формальное
решение.**



**Если выборка и приблизительная оценка
оказались довольно близко друг от друга...**

**Мы получили значение
вероятности, равное
5% или больше, когда
сравнили их...**

...что означает,
что наше среднее
выборочное значение
находится точно
внутри 95% площади
под куполом холма.



**...мы сделаем вывод, что
наша приблизительная
оценка, вероятно, была
сделана правильно.**

**Есть большая вероятность, что мы бы
произвольным образом сделали выборку,
похожую на нашу...**

...из генеральной совокупности,
центрированной прямо там.





Как бы то ни было, если выборка и предварительная оценка далеки друг от друга...

...при сравнении мы получили Р-значение меньше 5%...

...что значит, среднее значение в нашей выборке находится в «хвостиках»!



...мы можем и отказаться от них.

Очень маловероятно, что мы бы произвольно собрали выборку, подобную этой...

...из генеральной совокупности, центрированной прямо тут.

Поэтому я думаю, что настоящая генеральная совокупность не центрирована здесь.



Возможны лишь эти два варианта.

К сожалению,
ни одна из версий
не подходит
на 100%.



Потому что неважно, **каково будет наше формальное заключение...**

Это
оно?



Может
быть!

Наши данные
прекрасно
подходят под это
значение!



А может,
это оно?



Если это так, мы бы,
вероятно, не увидели
данных, похожих
на наши.

Поэтому,
думаю,
это не оно.



...есть вероятность, что мы заблуждаемся*!

Че-е-ерт!

Почему ты
не можешь
просто дать
мне ответ?



Извини, но наша
случайная выборка
могла оказаться
пустышкой!



* Это может показаться
довольно нудным, но все
это невероятно важные
вещи. Подробности
на стр. 222.

**Помните, проверка гипотез
основывается на оценке, построенной
с помощью одной выборки...**

*Мы немного погоняем
эту штучку
туда-сюда.*



**...и если вдруг окажется,
что наша выборка
неправильна...**

**...выводы тоже будут
неверны.**

*Если бы мы
случайно собрали
отклоняющуюся
от нормы выборку...*

*...наше
Р-значение
не имело бы
никакого
смысла!*



**К счастью, даже несмотря на то,
что мы никогда не можем добиться
абсолютной точности...**

*Никакое количество
тренировок
не может вам
гарантировать
100%-ного успеха.*



**...проверка гипотез может порой
оказаться очень полезной.**

Как мы увидим в следующей главе, все эти тонкости, возникающие при проверке гипотезы...

Оцени ситуацию!
Отмерь!

Толкай!

Переверни и толкай!

Высчитай Р-значение!

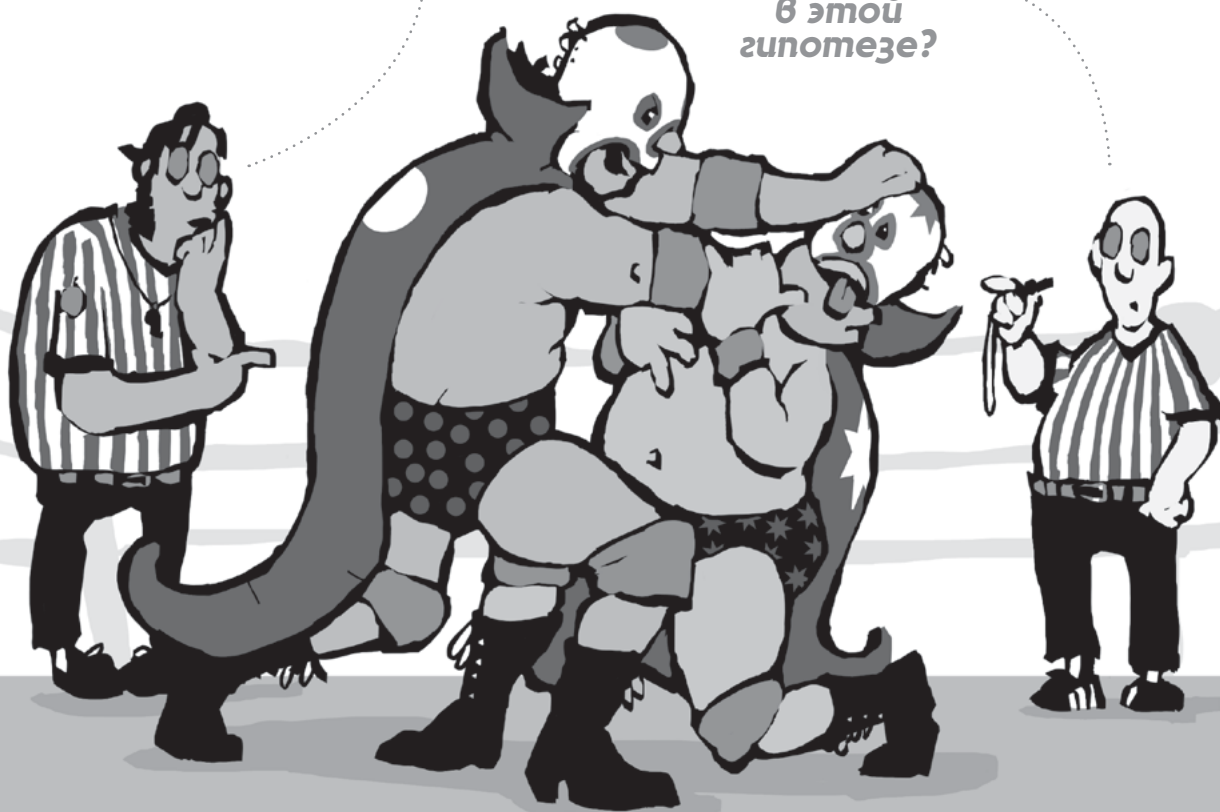
Прими решение!



...становятся особенно актуальными, когда мы пытаемся ответить на один конкретный вопрос:

будет ли доказательство достаточно сильным...

...чтобы заставить нас усомниться в этой гипотезе?



Глава 13

ПРОТИВОСТОЯНИЕ

Я лучше!



Теперь, изучив с формальной точки зрения, какие шаги предпринимают для проверки гипотезы...

...давайте посмотрим, как это происходит в жизни.

Оцени
ситуацию!
Толкай!



Просчитай!

**Прими
решение!**

Отлично, а теперь
отправляйся туда и покажи,
где раки зимуют!



На практике мы используем все эти шаги...

**...чтобы столкнуть
одну идею.....**

Я теперь совсем
другой!



...с другой.

Меня зовут Нолик...

...и я зануда.





Все это похоже
на реслинг...

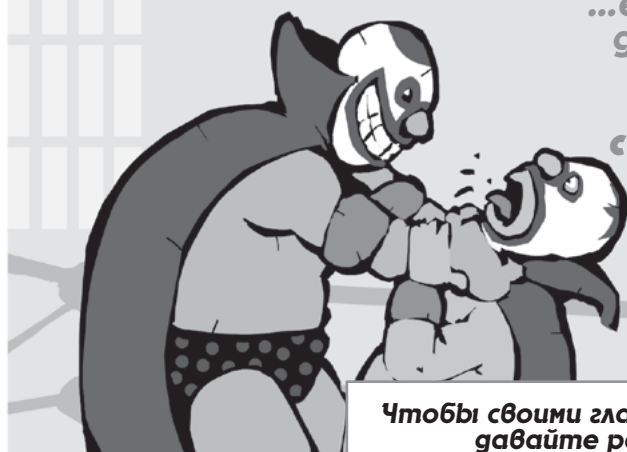


...но чтобы победить, здесь
все нужно делать наоборот.

Побеждает
всегда самая
неоригинальная
идея...



...если только наше
доказательство
не окажется
достаточно
сильным, чтобы
перевесить!



Чтобы своими глазами увидеть, как это работает,
давайте разберем несколько историй.

**Представьте себе,
что Минерва Хайтауэр
(назовем ее Доктор Счастье)...**

Привет!

Я самый главный
в мире специалист
по козням!

**...переживает из-за своей
установки, в которой
готовится яд.**

Она должна разливать в среднем
по 0,25 г чистого зла...

**...в каждую из этих склянок
с запатентованным балззамом,
вызывающим медленную смерть...**

Бульк!

Бульк!

Доктор Счастье
лично гарантирует
качество
продукции.

...но, кажется, установке пришел конец...

Хозяйка, клиенты
жалуются...

...некоторые
уверяют, что
в балззаме
недостаточно
зла...

Но, дорогой мой,
у меня репутация!
И я должна за ней
следить!

...а другие
говорят,
наоборот, его
слишком много.

Какое
расточительство!

**...поэтому
Минерва
начинает
проверку.**

**Итак, Доктор Счастье
делает случайную
выборку...**

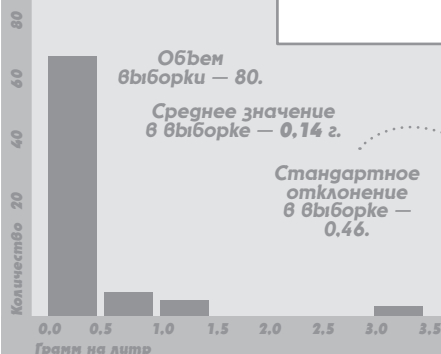
Надень-ка повязку, чтобы
не подсматривать, дорогая..

...и выбери
наугад
80 склянок.



...сортирует...

**...и, сделав подсчеты, узнает среднее значение:
0,14 г.**



Оказывается,
в каждой склянке
недостаточно
чистейшего зла.

**Должно быть,
сломалась моя
старушка!**



Но вот ведь проблема:

**Доктор Счастье
вознамерилась купить
новую установку!**

Давай просто
выкинем эту
рухлядь!

**Как же мне
хочется
купить
новенькую
ХТ-4300!**



**Но ее бухгалтер не дает это сделать,
пока они оба не убедятся,
что установку уже действительно
не спасти.**

Может быть, и правда
наша старая установка
работает хорошо...

...а наша выборка
показывает такие
результаты
действительно
случайно!



**Чтобы оценить достоверность,
Доктор Счастье может
воспользоваться проверкой гипотез.**

**И во время проверки Доктор Счастье
сталкивает между собой эти две идеи:**

либо эта старая,
повидавшая
виды установка
отжила свое...

...либо
нет.

Одна идея интересная...

...другая неоригинальная.

Я должна
купить новую
установку!

Только давайте
не будем спешить
и принимать
необдуманные
решения, Минерва.

**Каждую идею можно объяснить по-своему
и увидеть, почему мы получили данные,
которые получили...**

Среднее
значение
в выборке
показывает,
что жидкость
в склянке
какая-то
светлая...

...потому что
установка просто
больше неспособна
разливать
достаточное
количество чистого
зла!

Каждая отдельно взятая
выборка, вероятно, покажет,
либо недобор, либо перебор —
чисто случайно.

Что, если жидкость
в этой выборке
бледновата из-за
случайной вариации?

**...и теперь Доктору Счастью нужно посмотреть
на неоригинальное объяснение и подумать,
сможет ли она отклонить эту версию!**

Поэтому она берет свои данные и намечает предполагаемое распределение выборки.

Я строю ее на основе 80 случайно отобранных флаконов.

Затем она сдвигает распределение туда, куда велит неоригинальная идея...

Я перемещу его туда, где оно было бы центрировано, если бы установка не была сломана.

...и высчитывает Р-значение...

Если бы совокупность была по-прежнему центрирована по отметке 0,25...

...была бы всего 3%-ная вероятность...

...что я произвольно получила бы такое же среднее значение, как то, что вышло самым деле.

...затем берет время на обдумывание того, что же на самом деле означает полученное Р-значение.

Р-значение, которое меньше 5%...

...означает, что в долгосрочной перспективе мы бы случайно получили...

...меньше чем 1 из 20 средних значений в выборке, похожих на то, что я и получила...

...если бы машина по-прежнему выдвигала в среднем 0,25 г зла.

В итоге она со всей уверенностью отвергает неоригинальную идею...

...в угоду интересной.

Ох, милый, эта вероятность так мала...

...что вряд ли мы получили такой низкий средний показатель случайно.

Все же мне необходимо купить новую установку!

В этой истории Р-значение Доктора Счастье помогло ей принять уверенное решение.

Р-значение,
равное 0,03,
означает, что...

...я могу
быть на 97%
уверена
в этом!



Но помните, в статистике у уверенности всегда есть и обратная сторона.

Р-значение,
равное 0,03,
также означает,
что...

...в долгосрочной перспективе
мы в 3% случаев будем
получать такие же
неоднозначные выборки,
как и эта...

...а может, и сейчас
как раз так
и вышло.



Поэтому, даже если, кажется, есть доказательства в поддержку решения Доктора Счастье...

...она может ошибаться.

Я покупаю новую
установку, и ты меня
не остановишь!

А песенка этой
старой ржавой
развалины должна
быть спета!

Я просто говорю,
что она все еще
может работать,
не все так плохо.

Мы никогда
не можем быть
уверенными
в случайных
выборках.



К лучшему это или к худшему, но проверка гипотезы всегда заканчивается именно этим.

Мое решение
может быть
неверным!

К счастью,
в долгосрочной
перспективе это,
возможно, не так.

Ну-ка, дай мне мою
кредитку, дорогой,
я заказываю себе
новую установку!

Да не суетитесь
вы так!



**Как бы то ни было,
в конце концов Доктор Счастье
получила то, что хотела,
проверив свою гипотезу...**

**Сомневаюсь я
в обыкновенных,
неоригинальных
гипотезах!**



**Все, побежала
покупать себе
ХТ-4300!**

**У меня лучшая
работа
в мире!**

...но так бывает не всегда.

**А чтобы знать,
как оно бывает,
давайте рассмотрим
еще один случай.**



**Представьте себе,
что Безумный Билли
подкармливает червяков
в своем болоте стероидами...**

Это вещество
гарантирует улучшение
роста ваших червяков!

Но стоит
не очень-то
дешево!



...и хочет узнать,
эффективны ли они.

И вот он делает случайную выборку...



...28



...29



...30

...сортирует...



...и, сделав подсчеты,
получает среднее
значение, равное
4,19 см.



Хм-м, а результат
обнадеживает!

Конечно, он знает,
что среднее значение
в генеральной совокупности червей
раньше было 4 см...

Я помню каждую
банку, которую
когда-либо
продавал...

...а продал я
сто тысяч
миллионов!



...и надеется доказать,
что оно и правда стало больше.

Если это так...

...стероиды
делают свое
дело!



Если оно
не стало
больше...

...то меня надули!



Но вот ведь загвоздка:

у него на руках среднее значение,
которое больше прежнего...

Благодаря этой
выборке я думаю,
что стероиды
работают!



...но это, возможно,
только благодаря
случаю!

Может, я просто случайно
насобирал 30 аномально
глинистых червей...

...из генеральной
совокупности,
которая вовсе
не изменилась!



Чтобы решить, прав он или нет, Билли
решил проверить свою гипотезу.

И он сталкивает между собой эти две идеи:

**Либо моя затея
со стероидами
для червей
работает...**

...либо нет.

Одна идея новая и интересная...

**...а другая скучная,
неоригинальная.**

**Среднее значение
в генеральной
совокупности
на самом деле
поменялось!**

**Совокупность
как была,
так и осталась.**

**Каждая идея сопровождается своими
объяснениями, почему мы получили те данные,
которые получили...**

**Среднее значение в моей
выборке стало больше,
потому что среднее
значение в болоте стало
больше!**

**Среднее значение в моей
новой выборке стало больше,
потому что я совершенно
случайно собирал более
длинные червей.**

**...и теперь Билли нужно посмотреть
на неоригинальное объяснение и подумать,
сможет ли он его отклонить!**

И вот Безумный Билли берет свои данные и намечает предполагаемое распределение выборки.

Я сделал это, измерив 30 червей.



Затем он сдвигает распределение туда, куда велит неоригинальная гипотеза...



Я его передвинул на место того, что, как мне известно, было старым средним значением в генеральной совокупности.

...и высчитывает Р-значение.

Если бы генеральная совокупность была по-прежнему центрирована по отметке 4...

...была бы всего лишь 28%-ная уверенность...

...что я случайно получил бы такое же среднее значение, как то, что вышло в действительности.



В этом случае мы смотрим только на один затененный конец, потому что мы сфокусированы только на том, стало ли среднее значение в генеральной совокупности больше.

Но когда он понимает, что же на самом деле означает его Р-значение...

Р-значение, равное 28%...

...означает, что в долгосрочной перспективе мы бы видели данные, подобные моим, около 3 раз из 10...

...если бы настоящее среднее значение в генеральной совокупности было равно 4.

&%@#\$!



...он делает неутешительный вывод, что не может быть уверен в эффекте стероидов!

То, что в среднем червяки получились более длинными совершенно случайно, кажется абсолютно возможным.



**Конечно, заключение, которое
сделал Билли...**

О черт!

Я не могу
быть уверен
в том, что
мои стероиды
работают!

**...не означает, что его
стероиды не работают.**

Терпение,
возможно,
они все-таки
эффективны!

Ведь в среднем
твои червяки
оказались длиннее,
чем показало
предыдущее среднее
значение.

**Это всего лишь означает, что это доказательство
недостаточно веское, чтобы подтвердить
устраивающее его предположение.**

Ты мог получить
эти результаты
и наугад!

**Заключение не то чтобы
удовлетворительное...**

Ничего
интересного
тут
не происходит.

Мы вернулись
к тому, с чего
начали.

...тем не менее оно очень важное.

Мы рассмотрели два разных примера проверки гипотезы...

Я жаждала увидеть доказательство, что моя установка сломалась.

Я жаждал доказательств того, что черви стали длиннее.



В обеих историях красивая новая идея...

...и увидели два разных результата.

Я могу быть уверена в своем доказательстве. Оно верно.

А я нет.



...столкнулась со старой, скучной, самой обыкновенной.

Я такой сексуальный в этом эластичном костюме!

Ну, не знаю...



В обеих историях нам хотелось, чтобы новая идея оказалась правильной...

Вы же сами знаете, что желаете мне победы!

...при этом мы наделили неоригинальную идею преимуществом вызывать сомнения.

Но пока у вас не будет достаточно доказательств, чтобы меня опровергнуть...

...победителем буду я.



Существует веская причина, чтобы проводить проверки, подобные этой.

Весь смысл проверки гипотез в том,
чтобы убедиться,
что мы не делаем
преждевременных выводов.

**Я именно то,
что тебе
нужно, детка!**

Ей-богу,
он такой
неотразимый,
интересный,
просто
потрясающий!

Погоди!

Тебе нужно
убедиться,
что я не такой...

...прежде чем ты
отдашь своим
чувствам!





Глава 14



ЛЕТАЮЩИЕ СВИНЬИ, ПЛЮЮЩИЕСЯ ПРИЩЕЛЬЦЫ И ПЕТАРДЫ



Похоже,
будет буря!



**На протяжении нескольких
предвдущих глав мы учились
высчитывать доверительные
интервалы...**

**...и проверять
гипотезы.**

Я на 95% уверен,
что ты меня
не ненавидишь!

Я на 97% уверена...
...что моей установке
по производству зла
на самом деле пришел
конец!

**Как мы уже видели, обе стратегии
предполагают одни и те же основные шаги.**

Сначала мы берем
случайную
выборку...

...и используем, чтобы
представить себе
ее распределение.

Затем мы «отрезаем»
кусочки, находящиеся
на некотором отдалении
от середины, чтобы
высчитать вероятность...

...хотя иногда
получается
информативнее
сначала просто
передвинуть
выборку на новое
место.



Теперь, когда мы наконец разобрались
с самым основным...

Поздравляем!

Если вы понимаете,
как все это
работает...

...вы поняли
суть
статистики!

Остались
мелочи.



...мы собираемся посвятить эту главу разбору отдельных случаев...

...чтобы дать вам представление о том,
что вас ждет,
если вы захотите углубиться в тему.

Согласно прогнозу, ожидаются
летающие свиньи,
плюющиеся пришельцы
и петарды!

Нам, очевидно,
понадобится
это!



Пока что нас в основном интересовало,
каковы будут ваши действия в случае, если
УСЛОВИЯ ИДЕАЛЬНЫ:

мы научились охотиться
на среднее значение в одной
генеральной совокупности...

...используя одну
большую выборку...

Водичка
чистая...

Солнышко
светит...

...Статистика —
это легко!

...и по-настоящему
случайные замеры,
полученные
при ее изучении.

На практике же условия бывают гораздо
более неоднозначными...

А что, если не удастся
собрать выборку
достаточно большого
объема?

Что, если вы не сможете
сделать замеры,
которые будут
действительно
случайными?

А что если вас
интересует что-то
другое, а не среднее
значение?

...а продвинутая статистика во многом
предполагает умение решать сложные задачи.

Хорошая новость заключается в том, что неважно, насколько сильно задача усложнится со временем...

...(а так оно и будет)...

Отдам свое королевство за Р-значение!

Что-то я не очень уверен!

...мы все равно можем полагаться на основные шаги, о которых говорили на страницах этой книги...

...нам лишь нужно научиться их модифицировать.

Меняются детали...

...но базовые шаги остаются прежними.

ЛЕТАЮЩИЕ СВИНЬИ!

Вот наша первая история: давайте представим себе, что летающие свиньи в пятнышко быстрее, чем летающие свиньи в полосочку...



Подавись моей пылью, кекс!

...а еще они намного дороже...

И поэтому Сэм Нидлхаус хочет узнать, насколько они быстрее?

Я запускаю свой свинобизнес.

И мне интересно, инвестировать ли сбережения в пятнистых свинок...

...или лучше прикупить полосатых, а оставшиеся от покупки деньги пустить на симпатичные костюмчики для них?

Ты хочешь вложить деньги в скорость или тебе важнее стиль?

Чтобы принять это решение, тебе нужно понимание того, насколько именно пятнистые свиньи быстрее.



Если говорить языком статистики, вопрос звучит так:

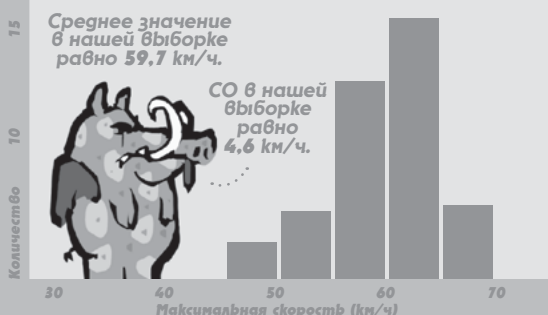
как мы наметим такой доверительный интервал...

...который нам расскажет о разнице между двумя средними значениями в генеральной совокупности?

Давайте соберем два комплекта среднестатистических выборочных данных и выясним это.

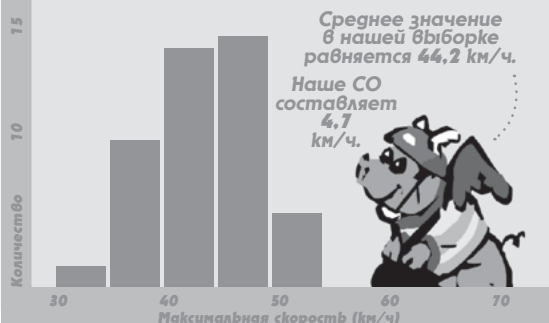


Тогда нам подойдут данные о 40 произвольно отобранных пятнистых свинках...



...чтобы наметить предполагаемое распределение выборки...

...и 40 произвольно отобранных полосатых свинках...



...которое немного отличается от привычного нам.

Мы центрируем его по разнице средних значений наших выборок...

15.5

...и высчитываем стандартное отклонение немного по-другому...

...но оно по-прежнему нормально распределено.

Это их лучшее предположение о разнице между средними значениями в выборках, если бы они произвольным образом насобирали сто тысяч миллионов выборок из каждой совокупности.



13.5

15.5

17.5



Да, теперь распределение выглядит немного иначе, но мы все равно можем его разделить на зоны...*

Так как распределение нормальное, мы можем отрезать «хвосты», в 2 СО от центра и сказать:

мы уверены на 95%...

...что в среднем пятнистые свинки на 13,5-17,5 км/ч быстрее полосатых.

...и использовать этот метод, чтобы принять уверенное решение.

Куплю, пожалуй, пятнистых!

Если они настолько быстрее...

...в долгосрочной перспективе это окупит все дополнительные вложения!



* На стр. 223 вы найдете формулу.

ПЛЮЮЩИЕСЯ ПРИШЕЛЬЦЫ!

У гигантских человекоядных жуков-пришельцев с планеты E8M-286 кислотная слюна...

Я бы не стал это использовать для увлажнения кожи лица!



Каков средний pH-фактор их слюны?



А она может прожечь бронезилет?



К сожалению, мы можем поймать и взять пробы только у **нескольких особей...**

Мы смогли получить интересующую нас информацию только о 10 случайно отобранных пришельцах...



...до того, как они проглотили нашего последнего специалиста по сбору данных.

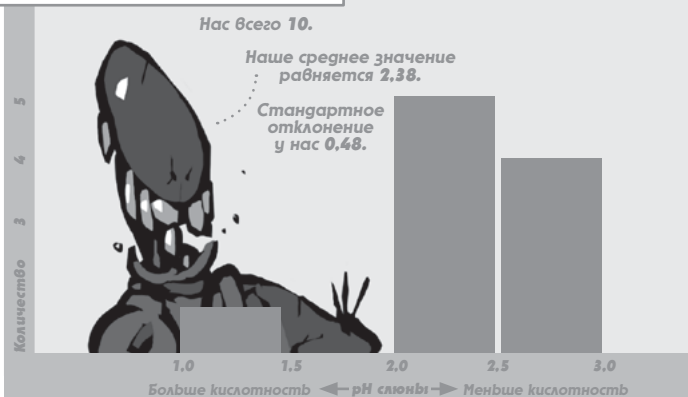
Но если объем выборки меньше 30, этого недостаточно...

...нам придется прибегнуть к другим техникам подсчета.



К счастью, есть одна замечательная стратегия, которой мы можем воспользоваться, если проблема лишь в **небольшом объеме выборки.**

Мы были бы только рады собрать больше данных, но не можем!



Но для начала нам нужно сделать одно вопиющее в своей дерзости предположение.

Если мы предположим, что генеральная совокупность нормально распределена...



Кажется, много природных явлений распределены нормально, может быть, тут будет так же.

...(что может быть весьма сомнительно)...

В таком случае сложно судить по 10 пришельцам.



...мы сможем использовать выборочные данные, чтобы построить предполагаемое распределение выборки...

...у которого будет немного более растянутая форма, нежели та, к которой мы привыкли.

Мы центрируем распределение по среднему значению в нашей выборке...

2.38

...и высчитываем стандартное отклонение все тем же старым способом...

Распределение выглядит очень похожим на нормальное, но в «хвостиках» у него больше вероятности...

...и оно оказывается распределенным не нормально!

Называется оно **t-распределением Стьюдента**.



2.04



2.38

2.72



Оно, конечно, более растянуто, но мы все равно можем высчитать степень достоверности...*

Для 95%-ного доверительного интервала в такого рода t-распределении мы отсчитываем от центра 2.26 СО вместо всего лишь 2.

Что означает, что мы уверены на 95%...

...что в среднем РН-фактор их слюны колеблется между 2.04 и 2.72.

А это что-то среднее между уксусом и лимонным соком!



...просто делать выводы нам нужно особенно осторожно.

Если наше предположение о нормально распределенной генеральной совокупности окажется неверным...

...нас может прожечь кислота.



* На стр. 223 вы найдете соответствующую формулу.

ПЕТАРДЫ!

Это ужасно, но маленькая Сбюзи Бикер хочет поиздеваться над соседской кошкой...



Эй, кус-кус-кус.

...подложив ей под хвост петарду!

Пш-ш-ш...



Она предпочитает петарды марки Dingalings...

...потому что, согласно надписи на упаковке, у них задержка срабатывания пять секунд...

Ба-бах!



МЯУ-У-У!

Идеальное время ожидания!



Но Сбюзи немного переживает, насколько они надежны.



Возможно, среднее значение и равно пяти секундам...

...но некоторые петарды взрываются быстрее, а некоторые — наоборот.

Мне не нравится, когда они взрываются у меня в руке...

...или когда приходится ждать так долго, что кошка успевает убежать!

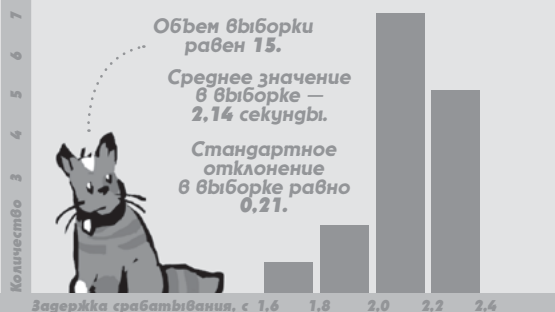


Пш-ш-ш...

Говоря языком статистики, вопрос Сбюзи сводится к вариативности...

...с этим можно разобраться, проанализировав одну выборку...

Меня волнует вопрос: когда я поджигаю петарду марки Dingaling, могу ли я рассчитывать, что она взорвется в течение пяти секунд, как сказано на упаковке, или хотя бы близко к этому?



...чтобы сделать предположение о размахе вариаций генеральной совокупности.

В этом случае нам лучше сделать предположение, основываясь на стандартном отклонении вместо среднего значения...

...и весь процесс потребует от нас совершенно других техник подсчета.

Осторожно! Огонь!

Пш-ш-ш...

Шаг назад!
Мы же специально обученные профессионалы!

Не пытайтесь повторить это дома!

Как бы то ни было, мы по-прежнему выполняем самые простые действия:

намечаем немного видоизмененное предполагаемое распределение выборки...*

Это самое смелое предположение о стандартном отклонении в выборке Dingaling, если бы мы в произвольном порядке собрали сто тысяч миллионов экземпляров!

...и отрезаем «хвостики» вероятности...

Так как же мы тогда узнаем, где именно отметит 95%-ный доверительный интервал?

Придется высчитывать на компьютере.

Или как высчитать Р-значение?

Оно не нормальное, оно смещено!



0.10

0.16

0.28

0.34

0.40



...чтобы сделать понятное всем заключение.

Основываясь на данных нашей выборки, мы уверены на 95%...

...что стандартное отклонение в генеральной совокупности Dingaling колеблется между 0,16 и 0,34 секундами...

...а если это правда, то вы можете рассчитывать, что большинство петард Dingalings будут взрываться в интервале между 1,5 и 2,4 секундами.



Ха-ха!

Пш-ш-ш...



Как следует из этих историй, нам нужно учитывать огромное количество нюансов...

...когда мы пытаемся ответить на вопросы продвинутой статистики.

А что, если сравнить плюющихся пришельцев и петарды?

Секундочку!

И, по правде говоря, нюансам нет числа.

Так, если придется иметь дело с данными, которые как-то друг с другом коррелируют...*

* См. стр. 224, если вас интересуют подробности.

Мы хотим знать среднюю температуру всех гекконов в этом тропическом лесу.

Но мы не можем сделать совершенно случайную выборку, потому что те гекконы, которые сидят на солнышке, теплее...

...чем те, что отбывают в тенишке.

Что означает, что наши температуры взаимосвязаны...

...так же, как и наши температуры.

...есть несколько хитростей, которые помогут нам в работе.

Если мы объединим всех гекконов коррелирующей структурой...

...мы сможем их использовать, чтобы прикинуть распределение выборки!

Если нас интересует характеристика, значение которой, кажется, **соответствует** значению другой...

Как сильно зависит скорость твоего уменьшения в размерах...

...от того, сколько уменьшающей в размерах жидкости ты пьешь?



...в нашем распоряжении есть совершенно другие хитрости.

Мы можем провести регрессионный анализ...

...что включает в себя проведение линии между двумя характеристиками на одной диаграмме...

...и оценку распределения выборки с помощью вычисления угла наклона линии.

Уменьшение (в см)

Доза (в граммах)



Несмотря на тот факт, что **продвинутая статистика** переполнена **разными приемами и хитростями...***

Не забудь о дисперсионном анализе...

...и как делать статистические выводы о пропорциях...

...и как предсказывать будущее!



...основные шаги формирования статистических выводов остаются теми же!

* На стр. 224-225 вы найдете более подробную информацию.

Всегда помните об этом, если захотите углубить свои знания в области статистики.

От нюансов поначалу может голова пойти кругом...

Если вы хотите научиться предсказывать погоду...

...вот тут целый мешок, полный информации об этом!



...но, по сути, все проблемы в статистике имеют схожую природу.

И выглядят они вот так:

как мы можем судить о генеральной совокупности...

...когда у нас есть доступ только к выборке?



Разбираться с ними мы будем следующим образом:

мы используем имеющиеся у нас данные, чтобы наметить распределение выборки...

...а затем «отрежем» хвостики вероятности...

...хотя иногда можно попробовать сначала передвинуть его на другое место.



Заключение

Мысль как статистик

Ом-м-м...



На страницах этой
книги, так уж вышло,
мы рыбачили...

Хотел бы
я переловить
всех пираний...

...но поймать
мы можем только
несколько
из них.

...занимались собирательством...

Вот тебе
защитный
шлем.

Он поможет
тебе избежать
субъективности.

...и охотились.

Мы никогда
не сможем
поймать его...

...но мы можем сделать
предположение
относительно
того, сколько их
группируется
вокруг.

Продельвая все это, мы учились
думать как статистики!

Мы не знаем
всего...

...но это
не значит,
что мы не знаем
ничего!

В первой части мы рассмотрели множества выборочных данных...

Вот засада!

Все смещено!

Ну-ка, расскажи нам об объеме, расположении и размахе вариаций.

Вот класс!

...и исследовали их.

Осторожнее с неизвестными переменными!

Затем, во второй части, мы изучали статистический вывод — ...

Что эти червячки...

...могут сказать нам о тех?

Все это было необходимо, чтобы высчитать вероятность!

...метод использования выборки для поиска характеристик в генеральной совокупности.

Мы научились намечать предполагаемое распределение выборки...

Вот одна банка.

А вот и план, как сто тысяч миллионов банок будут группироваться в долгосрочной перспективе.

Да это же центральная предельная теорема, класс!

Да вы все как с ума посходили!

...пристально взглядываться в него...

Э-эй, мы там?

...и, отрезав «лишнее», высчитывать доверительные интервалы...

Основываясь на данных нашей случайной выборки...

...мы уверены на 95%...

...что они вас совсем не ненавидят!

...или проверять гипотезы.

Хм, я практически уверен, что среднее значение в генеральной совокупности находится прямо тут.

А что мы тогда скажем на это?

**Наконец, мы научились
модифицировать эти основные
шаги...**



Бум!

**...когда вопросы становятся
более сложными.**

Когда меняются
обстоятельства...

...мы используем
распределение
выборки, имеющее
другую форму!



С тех пор как ученые
определились с **основными**
шагами по формированию
статистических выводов...

Вот это да!

Какая красивая
форма!



...эти шаги использовали самые разные люди...

Преступники
международного
уровня!



Ученые-
планетологи!



Пивовары!



Генералы!



...при самых разных обстоятельствах...

Я уверен на 68%,
что понимаю
этот секретный
шифр!



Я уверен на 95%,
что возраст
Вселенной колеблется
от 12 до 15
миллиардов лет.



Я уверен на 3%,
что вкус пива
из этого бочонка
изумителен...

...и оно
не так уж
сильно
дурманит.



Я на 99,7% уверен,
что война — дело
неблагодарное...

Но все равно
давайте
повоем!



И хотя вышеизложенное и есть суть статистики,
обладающей невероятной силой...

Мы можем использовать
статистические
методы всегда, когда
нам не хватает
информации...

...и принимать
уверенные
решения.

...она породила
немало загадок.



**За долгие годы терминология,
используемая в статистике,
расцвела пышным цветом...**

**...и лексикон
статистиков
серьезно пополнился...**

Нормальное
распределение...

...также
называется
распределением
Гаусса...

...а еще
t-распределением.

Хм...

**...поэтому научиться говорить
как статистики...**

**...может оказаться
непростым делом...**

Ой, полагаю, эти
явления не связаны
между собой!

Чего я ищу:
значимости или
значения?

Сколько
степеней
свободы
у остаточной
дисперсии?

А эта ошибка
считается
стандартной?

**...особенно, если вы
откроете самый конец
книги, где описаны
продвинутые методы.**

Повторяйте
за мной:

P-значение

F-значение

Z- или T-оценка

**Критерий
хи-квадрат**

И это всего
еще не все!

Я уже ни в чем
особо не уверен.

А вы уже думали
о непараметрическом
регрессионном
модулировании?

На протяжении многих страниц мы
наблюдали за тем,
как мыслят статистики.

Помните
о долгосрочной
перспективе...

...и не забывайте
про краткосрочную...

И все это
одновременно.

Да с этим
кто угодно
справится!

Но если вы хотите научиться
говорить как они...

...начните обучение...

Вам
понадобятся
перчатки...

...и каска.

Приложение Математическая пещера

Оставь надежду
всяк сюда
входящий...

...если только
вы не хотите
научиться
говорить
и писать...

...как
настоящий
статистик!

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\bar{x} \pm 2 \left(\frac{S}{\sqrt{n}} \right)$$

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



Когда при изучении выборки мы пишем формулы, то записываем все свои наблюдения таким образом:

СЛУЧАЙНАЯ ВЫБОРКА

Формирование выборки произвольным образом — принципиальный вопрос для статистики. Тут важно, чтобы она не отличалась систематически от генеральной совокупности, которую представляет.

Выборка, строго говоря, это собрание отдельных наблюдений за конкретными переменными (см. ниже). Выборка называется случайной, когда состоит из отобранных произвольным образом наблюдений, каждое из которых независимо от остальных.

Говоря о случайном отборе данных на страницах этой книги, мы, в частности, имеем в виду простую случайную выборку (ПСВ). Формально простая случайная выборка (ПСВ) размера n — это собрание n -ного количества наблюдений, полученных таким образом, что все возможные выборки n -наблюдений из генеральной совокупности имеют одинаковую вероятность быть отобранными.

Иногда срабатывают и другие неслучайные техники отбора, например систематический отбор или отбор по стратам (однородным частям в генеральной совокупности — прим. ред.), но, какую бы стратегию мы в итоге ни выбрали, мы должны быть уверены, что окончательная выборка будет представлять всю генеральную совокупность. Если этого не происходит, наши последующие действия оказываются бессмысленны.

$X_1, X_2, X_3 \dots X_n$

где X_1 — это первое наблюдение...

... X_2 — это второе наблюдение...

...а X_n — наше последнее наблюдение в списке, которое включает в себе n -ное количество наблюдений.



ОБЪЕМ ВЫБОРКИ (n)

Это общее количество экспериментов, собранных в одной выборке. В целом, чем больше выборка n , тем больше степень достоверности наших выводов. Но нужно следить за тем, чтобы выборка была сделана произвольным образом!

СРЕДНЕЕ ЗНАЧЕНИЕ В ВЫБОРКЕ (\bar{x})

Чтобы вычислить среднее значение в выборке, нужно сложить все показатели в ней и разделить их на ее объем. Вот по этой формуле:

Так-с...

...среднее значение в выборке называется «хбар»

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$



Среднее значение еще называется «средним арифметическим» или просто «средним». На страницах нашей книги мы избежали этого термина, отдавая предпочтение «среднему значению»: нам показалось, что, используя этот более привычный и понятный всем термин, мы облегчим понимание процесса формирования статистических выводов. А еще мы уверены, что большинство наших читателей, услышав слово «среднее арифметическое», все равно тут же подумают о «среднем значении».

Как бы вы его ни называли, среднее значение — самый важный показатель среднего значения распределения. Существуют и другие методы, которые могут помочь нам прояснить вопрос формирования определенной совокупности данных, — их выбор всегда зависит от ситуации.

Например, медиана — это «центральное значение» в выборке, и в случае смещения оно может быть предпочтительнее. Подобным образом, усеченное среднее значение вычисляется путем исключения небольшого процента самых маленьких и самых больших показателей; ему отдают предпочтение, когда в выборке есть экстремальные значения.

СТАНДАРТНОЕ ОТКЛОНЕНИЕ (S)

При подсчете стандартного отклонения нам нужно понять, каково среднее расстояние от среднего показателя. Вот как это сделать. Объясняем (буквально) на пальцах:

- 1) высчитайте расстояние между каждым замером x и средним значением в выборке \bar{x} . Это и называется отклонение;
- 2) возведите в квадрат каждое отклонение;
- 3) сложите все возведенные в квадрат отклонения;
- 4) разделите сумму на $n-1$ (остановившись здесь, мы получаем дисперсию);
- 5) извлеките квадратный корень из полученного значения.

Вот вам формула:



S =



1) высчитайте отклонение;

2) возведите в квадрат;

3) сложите все квадраты отклонений...

...пока не доберетесь до последнего.

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2$$

n - 1

4) Разделите на $n-1$.

5) Извлеките квадратный корень из полученного значения.

Обратите внимание, что мы делим на $n-1$, а не просто на n . И тому есть четкое математическое объяснение.

Технически, **вариативность** представляет квадратный корень из среднего арифметического всех квадратов разностей. А **стандартное отклонение** — это квадратный корень из дисперсии. Обратите внимание, что мы используем просто букву s , когда говорим о СО в нашей выборке.

ПЕРЕМЕННАЯ (X)

Переменная — то, что нас особенно интересует. Но так как в статистике мы всегда собираем данные произвольным образом, то называем те переменные, которые ищем, **случайными**. По определению, **случайная переменная** — это переменная, значение которой совершенно случайно.

В краткосрочном периоде мы не имеем никакой возможности предсказать значение случайной переменной, пока не получим ее саму (как, например, в случае подбрасывания монеты). В долгосрочном периоде мы предсказываем значение случайной переменной, используя **вероятность** (см. ниже).

Длина червей очень разнится.

И доход пиратов тоже.

Да и скорость графонов.



РАСПРЕДЕЛЕНИЕ

Говоря математическим языком, **распределение** описывает расположение всех возможных показателей случайной переменной. Если, например, бы создали гистограмму со всеми показателями переменной генеральной совокупности, у вас получится **распределение генеральной совокупности** для этой переменной.

Если говорить более общо, распределения позволяют нам высчитывать **вероятности** случайных показателей из конкретного интервала. Делая статистические **выводы**, мы высчитываем вероятности, используя **распределение выборки** (см. ниже). Но если бы мы имели на руках распределение генеральной совокупности, мы могли бы использовать и его для подсчета вероятности. Вот как это делается.

Если бы мы знали, каково распределение всей генеральной совокупности рыбы в озере, отсортированной по длине...

Отсюда следует, что, если бы мы нырнули в озеро и случайным образом **выбрали одну рыбешку**...

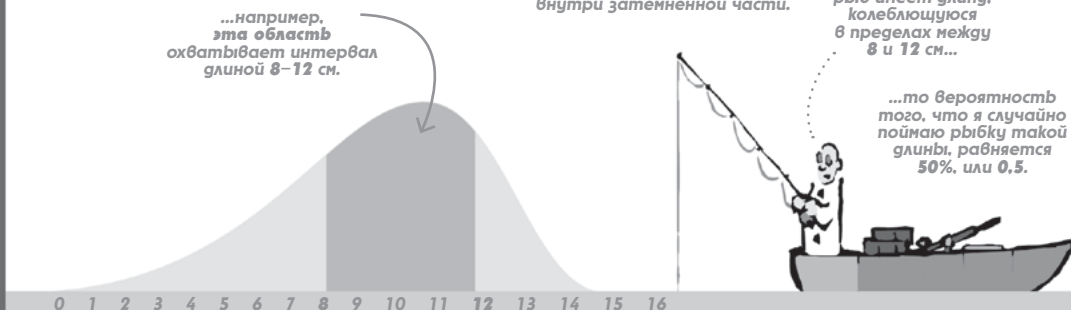
...мы могли бы подсчитать процентное соотношение рыбы в любой области этого распределения...

...вероятность того, что ее длина колебалась бы в пределах между 8 и 12 см, была бы такой же (в процентном соотношении), как и область распределения внутри затененной части.

Если половина всех рыб имеет длину, колеблющуюся в пределах между 8 и 12 см...

...например, эта область охватывает интервал длиной 8–12 см.

...то вероятность того, что я случайно поймал рыбу такой длины, равняется 50%, или 0,5.



Конечно, на деле мы никогда не можем посмотреть на распределение всей генеральной совокупности. Если бы мы могли, нам бы не была нужна статистика.

СТАТИСТИЧЕСКИЕ ДАННЫЕ ВЫБОРКИ ПРОТИВ ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Раз уж в статистике для суждения о генеральной совокупности используются выборки, для каждой из них есть свои термины и технические примечания.



Характеристики выборки называются **статистическими данными**.

Когда мы выводим формулы, **\bar{x}** подразумевает только среднее значение в выборке.

\bar{x}

s относится исключительно к стандартному отклонению в выборке.

s

Статистические данные — это то, что мы, собственно, изучаем, и потому судить о них можем со всей точностью.



Характеристики генеральной совокупности называются **параметрами**.

Греческая буква нижнего регистра **μ** относится исключительно к среднему значению в генеральной совокупности.

μ

Греческая буква нижнего регистра **σ** относится исключительно к стандартному отклонению в генеральной совокупности.

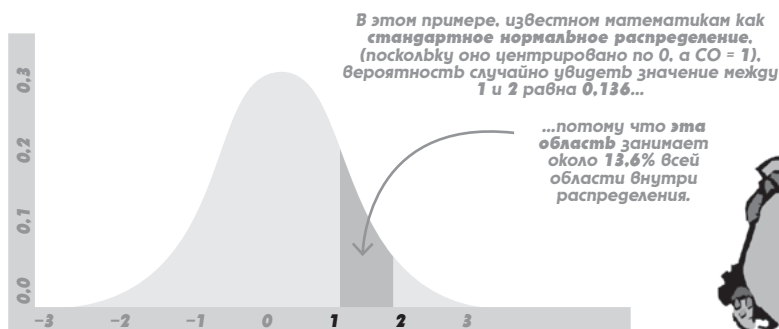
σ

Параметры — это то, что нас интересует в итоге, но о чем мы можем только строить предположения.

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

В математике и в теории вероятностей существует огромное количество распределений, которые бывают самых разных форм. Наиболее известно **нормальное**. В статистике оно особенно важно, потому что показывает, как группируются средние значения (см. «ЦПТ» ниже)

Как и в случае с любым другим распределением, мы всегда можем поделить нормальное распределение на области, которые будут отображать вероятности внутри него. Мы говорили о том, как это сделать, на стр. 115. Но вот еще пример.



РАСПРЕДЕЛЕНИЕ ВЫБОРКИ

Говоря технически, это распределение статистической выборки. И хотя мы можем наметить распределение выборки для любого вида статистики (СО, медианы и т. д.), мы здесь фокусируемся на распределении выборки, сделанном на базе средних значений. Так, например, если бы мы отобрали из генеральной совокупности много-много выборок размера n , высчитали бы \bar{x} для каждой из них, затем нарисовали гистограмму всех \bar{x} значений, то получили бы распределение выборки \bar{x} . Примером может послужить сарайчик Безумного Билли, где он хранит снасти (см. стр. 107). Распределение выборки послужит ключом к статистическим выводам.

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА (ЦПТ)

Многие статистические выводы зависят от **ЦПТ**, которая утверждает, что распределение выборки \bar{x} становится приблизительно нормальным, по мере того как увеличивается объем выборки n .

Если говорить более детально, то для **случайных выборок** большого объема n , отобранных из одной генеральной совокупности со средним значением μ и СО σ , распределение \bar{x} будет приблизительно нормальным со средним значением μ и СО, равным σ , деленной на **корень квадратный из n** .

Вне зависимости от формы, если генеральная совокупность содержит эти показатели...

μ

...распределение всех возможных средних значений в выборке большого объема n , случайно отобранной из той генеральной совокупности, будет содержать эти показатели и будет вот такой нормальной формы:

μ

σ/\sqrt{n}

Это распределение выборки \bar{x}



распределение в генеральной совокупности

распределение всех возможных показателей \bar{x}

σ/\sqrt{n}

также известна как стандартная ошибка.

ЦЕНТРАЛЬНАЯ ПРЕДЕЛЬНАЯ ТЕОРЕМА (ПРОДОЛЖЕНИЕ)

ЦПТ — это очень общий результат, который почти всегда будет применяться так, как описано в этой книге. Однако есть несколько **важных условий**, лежащих в основе ЦПТ.

Во-первых, ЦПТ работает только в том случае, если все показатели $x_1, x_2, x_3, \dots, x_n$ в нашей выборке взяты из **того же самого распределения генеральной совокупности**. Это обычно касается выборок, полученных на практике, но может быть важно, если мы разбираемся и с более сложными вопросами.

Во-вторых, каждый замер x_i должен быть **случайным**. Строго говоря, это означает, что все показатели x_i должны быть **независимы друг от друга**. Например, замеры температур, сделанные в каком-то географическом регионе, не будут независимы, поскольку результаты измерений из одного конкретного места будут очень похожи на замеры температуры в другом месте поблизости. Статистики скажут, что они «**коррелируют друг с другом**», потому что существует некий общий принцип, оказывающий влияние на значение каждого из x_i (см. «Корреляция» ниже).

Наконец, ЦПТ применима, когда n стремится к бесконечности, однако на практике мы используем приближенную версию ЦПТ, которая работает, когда $n \geq 30$. Мы называем такие значения n «**большими**». Это звучит довольно условно, но, если вдаваться в подробности, мы погрязнем в математике.

См. стр. 112

ВЕРОЯТНОСТИ

В книге мы обозначаем вероятности процентами (например, 95%), но в математике для тех же целей используют числа от 0 до 1 (например, 95% = 0,95). Так, формально **вероятность** — это число между 0 и 1, которое отображает степень возможности наступления какого-то события. Однако вся сложность заключается в том, что ее **высчитывают только в долгосрочной перспективе**.

Если, например, среди избирателей одинаковое количество мужчин и женщин, вероятность того, что случайно отобранный избиратель окажется женщиной, будет равна 0,5. Но первые несколько избирателей, отобранных произвольным образом, вполне могут оказаться мужчинами чисто случайно. То есть показатель 0,5 относится к **долгосрочной перспективе**: если мы произвольным образом отберем достаточно избирателей, в **конечном счете** получим равное количество мужчин и женщин.

В другом примере, когда мы подбрасываем монетку, с вероятностью 0,5 выпадет либо орел, либо решка. Но если мы подбросили монетку только один раз и выпал орел, вероятность того, что в следующий раз будет то же, равна **по-прежнему 0,5**. В этом случае каждый бросок монетки **независим** от другого.

Подводя итог, скажем, что всякий раз, когда мы **высчитываем вероятность**, она может быть выражена числом между 0 и 1 (или 0 и 100%), и это число всегда соотносится с областью внутри вероятностного распределения. По определению, вся область внутри него равна 1.



ВЫЧИСЛЕНИЕ ВЕРОЯТНОСТИ

Мы можем вычислить области внутри любого распределения (например, нормального, как мы видели на картинке на стр. 114), используя интегрирование, которое представляет собой **вычислительный метод**. На практике все необходимые вычисления делаются на компьютере.

В примере с сарайчиком Билли распределение выборки имеет нормальную форму из-за ЦПТ. Однако в огромном количестве других сфер применения статистики какое-то конкретное распределение выборки может и не иметь нормальной формы, но мы по-прежнему сможем сделать все необходимые вычисления.

Если хотите подробностей, вернитесь к главе 14.

Все распределения можно изобразить в качестве кривых, но также они могут быть представлены функциями. Их принцип похож на принцип работы компьютера, где есть **исходные данные** (в нашем случае случайные переменные) и **результат** (в нашем случае **вероятность**).

Среди условных знаков есть обозначение, принятое для функции вероятности f со средним значением μ и σ :

Если X представляет собой дискретную случайную переменную с распределением f , зависящим от μ и σ ...

...то f , зависящее от μ и σ (x), равно вероятности того, что X примет значение x .

К сожалению, дальше все только еще больше усложняется и запутывается. Вот, например, функция вероятности для нормального распределения:

$$h_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Хотя на первый взгляд она выглядит устрашающе, в целом функции вероятности, подобные этой, оказываются необычайно полезными, потому что они связывают определенные виды случайных событий (например, попытки понять рыбу определенного размера) с предсказуемыми долгосрочными выводами (как часто вы сможете рассчитывать на это в долгосрочной перспективе).



ОЦЕНКА РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

На практике, когда мы используем ЦПТ, мы никак не можем узнать настоящие значения параметров μ и σ , поэтому используем статистику \bar{x} и S , чтобы приблизить их друг к другу. Аппроксимация работает потому, что мы собираем данные произвольным образом. В результате мы ожидаем, что \bar{x} будет отличаться от μ , а S — от σ , но только из-за случайной вариации.

После того как мы заменили приближенные значения, мы можем называть получившийся результат **предполагаемым распределением выборки**.

На стр. 217 изображено настоящее распределение выборки...

...и намечаем мы его с помощью вот этого.

Это называется **предполагаемое распределение выборки**.



Предполагаемое распределение всех возможных значений \bar{x}

Обратите внимание, что мы можем построить **предполагаемое распределение выборки** и для других статистических данных, например S (см. стр. 201, история про петарды). Однако мы можем ждать, что распределение выборки будет нормальной формы, только если применим ЦПТ или похожие результаты.

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Строго говоря, это такой тип интервальной оценки, который отражает определенную степень достоверности. Доверительный интервал можно высчитать для любого параметра, хотя какие-то специфические технические детали могут меняться. Вот формула, по которой можно высчитать девяностопятипроцентный доверительный интервал для среднего значения в генеральной совокупности μ :

$$\bar{x} \pm 2 \left(\frac{s}{\sqrt{n}} \right)$$

Мы используем \bar{x} , чтобы высчитать среднее значение в генеральной совокупности.

Про эту формулу мы говорим: «среднее значение плюс-минус два стандартных отклонения».

Вот как высчитывается CO из \bar{x}

А плюс и минус означают, что мы отходим от середины в обоих направлениях.

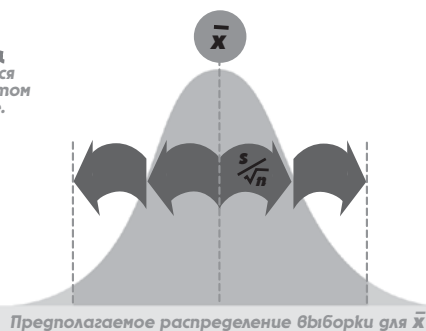
Мы называем это «отсечкой». Она говорит о том, насколько далеко в «хвосты» можно отойти в распространении, чтобы получить вероятность той степени, которую мы хотим.

Вот какое заключение мы можем сделать из этой формулы.

Мы на 95% уверены...



...что μ находится где-то в этом пределе.



Мы можем изменить степень достоверности, передвинув отсечку. Например, если бы нам был нужен доверительный интервал в 80% для среднего значения в генеральной совокупности, наша отсечка была бы на отметке 1,3, потому что приблизительно 80% нормального распределения имеет 1,3 CO относительно центра (см. пример на стр. 157).

В идеале нам нужен максимально узкий интервал для любой степени достоверности, поскольку он точнее. Одним из безошибочных способов сузить интервал будет увеличение n (для этого нужно собрать больше материала для наблюдений). Вот почему чем больше выборка, тем лучше! (см. пример на стр. 159).

Помните, что степень достоверности зависит от значения вероятности. Поэтому все это имеет смысл, только когда мы думаем о долгосрочной перспективе. В результате, высчитывая интервал с помощью формулы, приведенной выше, мы даже не знаем, учитывает ли она μ или нет! Все, что мы можем сказать, это то, что интервалы, намеченные таким образом, будут точными в долгосрочной перспективе. В случае с 95%-ным доверительным интервалом мы понимаем, что наши заблуждения составят 5% случаев... в долгосрочной перспективе.



ПРОВЕРКА ГИПОТЕЗЫ



При проверке гипотез используются те же самые статистические методы, что и при вычислении доверительных интервалов. Начинаем мы с построения предполагаемого распределения выборки. Но на этот раз оно нам нужно, чтобы ответить на вопрос: истинно или нет какое-то конкретное значение для параметра генеральной совокупности. И делаем мы это, проверяя, насколько совместимы исследуемые нами данные с тем конкретным значением.

Формально проверка начинается с рассмотрения двух гипотез: нашей исследуемой (иногда ее называют *альтернативной*) и *нулевой* (в книге мы использовали слова «скучная», «неоригинальная», «обыкновенная»).

Проверка гипотезы заканчивается тем, что мы вычисляем *P-значение* и используем его, чтобы вынести *формальное решение*, находятся ли наши данные достаточно далеко от параметра, предсказанного нулевой гипотезой, чтобы оправдать выбор, сделанный в пользу иного объяснения.

Вот краткий комментарий к описанному выше.

Наша нулевая гипотеза сильно сократилась.

Если μ на самом деле находится прямо здесь...

μ

...то в долгосрочной перспективе мы вряд ли...

Поэтому, если \bar{x} , который мы на самом деле нашли, находится в «хвостиках», а *P-значение* меньше 0,05, то нулевая гипотеза может быть ошибочной.

Хм-м-м...

...сможем в произвольном порядке найти значения для \bar{x} в этих «хвостиках».



Предполагаемое распределение выборки для \bar{x} , при условии, что нулевая гипотеза верна

В этой книге мы проверяем гипотезы, касающиеся средних значений. На практике эти же общие принципы применимы к любому параметру и соответствующей статистике — отличатся будут только математические данные.

P-ЗНАЧЕНИЕ

Формально *P-значение* — это вероятность ошибки при отклонении нулевой гипотезы. В книге мы обозначали *P-значение* процентами, хотя часто бывает и так, что его обозначают числами 0–1. *P-значение*, равное 5%, часто выражается числом 0,5.

Иногда мы вычисляем *P-значение* для обоих «хвостиков» нашего предполагаемого распределения выборки (см. стр. 157, это называется «двусторонний тест»), а иногда мы вычисляем *P-значение* только для одного из них (см. стр. 159, это называется «односторонний тест»). Какой из них выбрать, зависит от того, какого рода исследуемая гипотеза нас интересует.

На практике для вычисления *P-значения* пользуются компьютерами. Важно отметить, что (когда мы проводим двусторонний тест) вероятность 0,5 — это именно та область, которая не помещается внутри 95%-ного доверительного интервала. Можно прибегнуть к сравнительно простому способу проверки гипотезы: наметить 95%-ный доверительный интервал для μ , как описано выше. Если значение μ , предсказанное нулевой гипотезой, не находится в этом интервале, значит, *P-значение* меньше 0,05.

В подобных случаях увеличение n приводит к уменьшению *P-значения*. Поэтому, по мнению статистиков, сбор большего количества материала для исследования — это безошибочный способ с полным на то правом отклонить нулевую гипотезу. Это еще раз доказывает, что чем больше объем выборки, тем лучше!



Р-ЗНАЧЕНИЯ (ПРОДОЛЖЕНИЕ)

Помните, что Р-значение измеряет вероятность, поэтому оно актуально, только когда мы думаем о долгосрочной перспективе.

На практике мы отклоняем нулевую гипотезу, если наше Р-значение «достаточно мало», что (согласно общепризнанному правилу) означает меньше 0,05. Но в этом числе нет ничего магического. Вероятность «меньше чем 0,05» означает то же, что и «меньше 1 из каждого 20 случаев в долгосрочной перспективе».

Так, например, если мы проверяем гипотезу и у нас получается Р-значение, равное 0,049, это означает, что «если нулевая гипотеза была бы истинной, мы бы чисто случайно видели данные, подобные нашим, примерно 49 раз из каждой 1000 в долгосрочной перспективе». Так как 49/1000 меньше, чем 1/20, мы бы решили, что наши данные не очень хорошо соответствуют нулевой гипотезе.



См. стр. 172

МЫ МОЖЕМ ОШИБАТЬСЯ

В основе всех статистических исследований лежит случайная выборка, а в основе всех статистических выводов лежит вычисление вероятности. В результате всякий раз, когда мы используем статистические данные малых выборок, чтобы сделать предположение о каком-нибудь параметре в генеральной совокупности, мы можем ошибаться!

Из-за этого нам нужно очень внимательно следить за языком, ведь нас так и подмывает порой сделать громкое заявление, опираясь на статистические выводы. Нам нужно быть особенно осторожными, когда мы делаем формальные выводы, основываясь на Р-значениях. Потому что мы используем Р-значения только тогда, когда исследуем теории, которые нам особенно по душе.

Если мы проверяем теорию и используем маленькое Р-значение, чтобы подтвердить ее, есть вероятность, что мы заблуждаемся. Наша теория может оказаться ошибочной, и тогда наши результаты лучше объяснить случайной вариацией. В статистике это называется ошибкой первого рода.

Если же мы проверяем теорию с использованием больших Р-значений с целью отклонить ее, мы можем ошибаться, а наша теория — быть верной, и мы получим результаты, близкие по значению к нашей нулевой гипотезе. В статистике это называется ошибкой второго рода.

Резюмируя, скажем, что проверка гипотезы сводится к поиску ответа на вопрос «Какова вероятность того, что мы получили результаты совершенно случайно?». Их ведь нельзя использовать ни для убедительного опровержения, ни для убедительного доказательства теории. Они нужны только для того, чтобы помочь нам дискредитировать нулевую гипотезу.

В статистике, так уже повелось, мы в любую секунду можем оказаться неправы. И так оно всегда и бывает. А все потому, что мы привязаны к долгосрочности, когда оцениваем краткосрочные наблюдения.



СТАТИСТИЧЕСКИЕ ВЫВОДЫ О РАЗНИЦЕ



Чтобы высчитать доверительный интервал, касающийся разницы между двумя средними значениями в генеральной совокупности, мы можем использовать формулу, которая совсем немного отличается от той, что мы описали выше.

В этом случае нас интересует разница между двумя средними значениями генеральной совокупности, и мы высчитываем ее с помощью двух средних значений в выборке.

Вот так мы объединяем вариации двух генеральных совокупностей.

$$(\bar{X}_1 - \bar{X}_2) \pm 2 \left(\sqrt{\frac{S_1^2 + S_2^2}{n}} \right)$$



Плюс и минус означают, что мы сдвинулись в сторону от середины в обоих направлениях.

Мы используем отсечку, представленную цифрой 2, если хотим 95%-ный доверительный интервал. Но, конечно, допустимы любые изменения.

Все это — со нашего предполагаемого распределения выборки. В этом случае оно равно 1.

В нашем случае $\bar{X}_1 = 59,7$, $S_1 = 4,6$, $\bar{X}_2 = 44,2$, а $S_2 = 4,7$.

Необходимо отметить, что эта формула может иметь другой вид. Например, она изменится, если мы используем разные объемы выборки или если они слишком маленькие, чтобы получилось нормальное распределение.

СТАТИСТИЧЕСКИЙ ВЫВОД О ВЫБОРКЕ НЕБОЛЬШОГО ОБЪЕМА

Когда бы мы ни формировали статистические выводы о среднем значении генеральной совокупности при наличии выборки маленького объема (например, когда n меньше 30), мы не можем положиться на ЦПТ. В таком случае нам понадобится то, что известно как t -распределение. Оно работает, только когда сама генеральная совокупность нормальна.

По каким-то не вполне понятным историческим причинам, связанным с тем фактом, что этот вид распределения был открыт и впервые описан парнем, работавшим на пивоварне Guinness, t -распределение еще называют **распределением Стьюдента**.

Оно шире нормального распределения и на практике видоизменяется в зависимости от n (чем меньше n , тем шире t). В результате, когда мы используем t -распределение вместо нормального, наши доверительные пределы становятся шире, а наше P -значение — больше. В обоих случаях нам требуются дополнительные статистические данные, чтобы добиться той же степени достоверности. Как будто нас наказывают за небольшой объем выборки.

В нашем случае специфическое распределение, которое мы использовали, и представляет собой пример t , которое называют « t с 9 [как в случае с $n - 1$] степенями свободы».

А вот и замеры уровня РН, которые мы делали у 10 произвольным образом отобранных пришельцев: 2,09; 2,39; 1,32; 2,99; 2,62; 2,60; 2,45; 2,13; 2,27 и 2,95. Имея эти данные, мы можем высчитать, что $\bar{x} = 2,38$, $S = 0,48$, а $n = 10$. Затем мы вставляем все эти значения в следующую формулу и получаем 95%-ный доверительный интервал.

t -распределение выглядят почти нормально. Единственное, они немного шире вот в этих частях:

$$\bar{x} \pm 2,26 \left(\frac{s}{\sqrt{n}} \right)$$

Мы используем другую отсечку в зависимости от того, какого рода у нас t -распределение, и от того, насколько хотим быть уверены.



СТАТИСТИЧЕСКИЕ ВЫВОДЫ О СТАНДАРТНОМ ОТКЛОНЕНИИ



В истории с маленькой хулиганкой Сьюзи мы не можем рассчитывать на ЦПГ. Распределение выборки со стандартным отклонением не обязательно будет нормальным, и при вычислении доверительного интервала мы не можем пользоваться ничем, похожим на формулу, приведенную на стр. 220. Но основные шаги по-прежнему те же: мы создаем определенный вид распределения выборки (при помощи каких-нибудь эффективных математических вычислений), намечаем 95%-ный доверительный интервал и вычисляем все вероятности в этом пределе (опять же, пустившись в сложные подсчеты).

Вся эта эффективная математика слишком запутанна, чтобы объяснять ее здесь, но в основе всех вычислений все равно лежат показатели задержки срабатывания 15 случайным образом отобранных петард Dingalings: 2,05; 2,25; 2,33; 2,40; 1,66; 2,39; 1,89; 2,18; 2,06; 1,89; 2,14; 2,38; 2,07. Проанализировав эти данные, получаем $\bar{x} = 2,14$, $S = 0,21$ и $n = 15$.

Обратите внимание: чтобы сформулировать статистические выводы, нам нужно предположить, что генеральная совокупность нормальна. В этом случае предположение может быть справедливым, но если мы имеем дело с каким-то дефектом производства, который смещает всю генеральную совокупность (например, неожиданно большая задержка срабатывания у петард с одной фабрики может быть сбоем, а при анализе этого не учтут), тогда наши выводы могут оказаться ошибочными.

Так же, как и t -распределение, распределение выборки со стандартным отклонением меняется в зависимости от объема выборки. И, как и во всех случаях, чем больше n , тем более мы уверены в наших выводах.

КОРРЕЛЯЦИЯ

Корреляция искажает огромное количество статистических анализов. Если замеры нашей выборки $x_1, x_2, x_3 \dots x_n$ «коррелируют» друг с другом, они не будут независимыми и мы не можем применить к ним ни одну из наших методик. Поэтому в случае, когда данные коррелируют друг с другом, при формировании статистических выводов нам необходимо принимать этот факт во внимание.

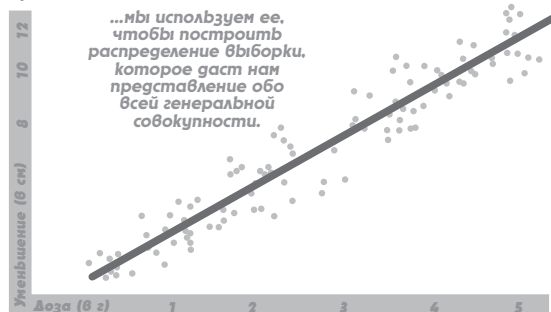
Например, при исследовании состояния здоровья двух биологических близнецов следует учитывать тот факт, что все данные об одном из них непременно коррелируют с данными брата или сестры, но больше ни с чьими. Эту корреляцию можно уменьшить, применив «парный тест». Другие исследования предполагают замеры, коррелирующие друг с другом по географическому принципу (как было в примере с гелконами на стр. 202) или во времени (представьте себе случай, когда PH-уровень слюны одного пришельца варьируется в зависимости от времени суток и за отчетный период мы получаем огромное количество замеров у одного и того же инопланетянина). Иногда мы можем внедрить этот тип корреляции, превосходящий его в более крупных математических моделях. Хотя, строго говоря, каждый вид корреляции предполагает применение своих трюков.



РЕГРЕССИОННЫЙ АНАЛИЗ

Он необходим, когда нужно исследовать природу взаимоотношений зависимой переменной и одной или нескольких независимых переменных. В примере на стр. 203 лекарство, заставляющее выпившего уменьшиться в размерах, будет независимой переменной, а то, насколько он уменьшается, — переменной зависимой.

После того как мы найдем идеально вписывающуюся в наш график линию...



В этом примере мы на 95% уверены, что увеличение дозы на 1 грамм...

...влечет за собой уменьшение роста на 2,3–2,5 дополнительных сантиметров.

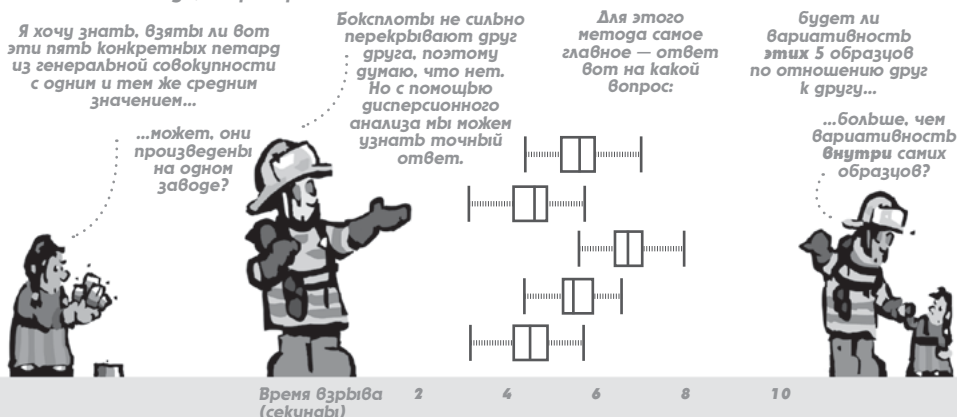


На практике нам нужно быть предельно внимательными, когда мы проявляем любопытство относительно того, могут ли независимые переменные на самом деле привести к изменениям в переменных зависимых. Причинно-следственная связь — вещь очень тонкая.

Мы также можем использовать этот тип анализа, чтобы сравнивать характеристики, например вес или рост, в генеральной совокупности. В таком случае наклон прямой скажет нам об изменениях в весе, происходящих по мере увеличения роста на каждый сантиметр.

ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионный анализ представляет собой технику проверки гипотез, но она сильно отличается от того способа, который мы рассматривали на страницах этой книги. Эта техника построена на сравнении вариантов между группами, а также внутри них. Существует множество способов применения этого метода, например:



СТАТИСТИЧЕСКИЕ ВЫВОДЫ О ПРОПОРЦИЯХ



Если нам интересно, **какова пропорция** футбольных фанатов, предпочитающих свинные шариканые ребрышки, и **каков процент** избирателей, которые с большой долей вероятности отдадут голоса за переизбрание сенатора Сэма Ворна на предстоящих выборах, мы можем использовать методы формирования статистических выводов, о которых говорили в этой книге, но нам нужно уточнить детали.

Например, таким образом мы можем высчитать 95%-ный доверительный интервал для пропорции совокупности p .

Мы берем нашу пропорцию выборки \hat{p} , чтобы оценить пропорцию генеральной совокупности.

Имея выборку большого объема, мы можем использовать стандартное нормальное распределение. Итак, чтобы получить вероятность, равную 95%, мы используем 2 отсечки.

$$\hat{p} \pm 2 \left(\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

Это стандартное отклонение пропорции выборки.

Еще раз, нам нужна выборка большого объема, чтобы все получилось, и чем больше будет объем, тем уже будет наш интервал!

В опросах это называется предел погрешности.

ПРЕДСКАЗАНИЕ БУДУЩЕГО

В нашей книге мы фокусировались на использовании данных выборки для того, чтобы проанализировать генеральную совокупность, ее среднее значение или стандартное отклонение. Но мы также можем использовать статистические выводы, чтобы сделать предсказания о каких-то отдельных наблюдениях. Например, мы можем задавать себе вопросы типа этого: «Основываясь на имеющихся у меня замерах, можно предположить, каким, вероятнее всего, будет следующий замер (x_{n+1})?»

Опять-таки, мы можем продвинуться немного вперед, предприняв основные шаги, необходимые при формировании статистических выводов. Например, если предположить, что генеральная совокупность нормальная (всегда опасное предположение), мы можем создать «интервал прогноза», который будет похож на стандартный доверительный интервал, но окажется немного шире.

На практике это используется, чтобы предсказать погоду или цены на финансовых рынках, хотя там применяют более сложные техники.





Об авторах

Грейди Клейн — художник-мультипликатор, карикатурист и иллюстратор, соавтор книг «Микроэкономика. Краткий курс в комиксах» и «Макроэкономика. Краткий курс в комиксах», а также создатель серии графических новелл «Затерянная колония». Живет в Принстоне (США) с женой и двумя детьми.

Алан Дебни — доктор философии, отмеченный многочисленными наградами адъюнкт-профессор статистики в Техасском Университете А&М. Живет в Колледж-Стейшен (США) с женой и тремя детьми.

Максимально полезные книги от издательства «Манн, Иванов и Фербер»

Заходите в гости:

<http://www.mann-ivanov-ferber.ru/>

Наш блог:

<http://blog.mann-ivanov-ferber.ru/>

Мы в Facebook:

<http://www.facebook.com/mifbooks>

Мы ВКонтакте:

<http://vk.com/mifbooks>

Предложите нам книгу:

<http://www.mann-ivanov-ferber.ru/about/predlozite-nam-knigu/>

Ищем правильных коллег:

<http://www.mann-ivanov-ferber.ru/about/job/>

Научно-популярное издание

The Cartoon Introduction to Statistics

By Grady Klein and Alan Dabney, Ph. D.

Грейди Клейн
Алан Дебни

Статистика
Базовый курс в комиксах

Главный редактор *Артем Степанов*
Ответственный редактор *Ольга Киселева*
Научный редактор *Ирина Николаева*
Литературный редактор *Анна Санникова*
Арт-директор *Алексей Богомолов*
Дизайн обложки *Сергей Хозин*
Верстка *Людмила Гроздова*
Корректоры *Мария Кантурова, Надежда Болотина*

НАУКА В СЛОВАХ И КАРТИНКАХ

Грейди Клейн — графический дизайнер, аниматор и художник комиксов из Принстона. Автор графических романов и соавтор нескольких научно-популярных комиксов.

Алан Дебни — профессор статистики Техасского университета A&M, исследователь больших данных, лауреат нескольких профессиональных премий за креативные решения в преподавании.

Комикс с драконами, великанами и инопланетянами в главных ролях — на самом деле неожиданно захватывающее и доступное пособие по статистике. Его создатели — профессор статистики Алан Дебни и художник Грейди Клейн — нескучно объясняют, как собирать надежные данные, делать правильные выводы, владея ограниченной информацией, оценивать результаты опросов и обращаться с множеством цифр, окружающих нас. Вы получите базовые знания о сложной науке самым увлекательным способом. А если, пройдя этот квест, захотите углубиться в изучение статистики, в конце книги вас ждет «Математическая пещера», полная формул и пояснений. Эта книга — пожалуй, лучший путеводитель по миру, которым правят данные.

* * *

Эта книга — лекарство для всех, кто страдает фобией статистики.

Чарльз Уилан, автор книги «Голая статистика.
Самая интересная книга о самой скучной науке»

Слава богу, кто-то умудрился написать книгу о статистике, которую весело читать. И будьте осторожны: возможно, вы не захотите с ней расстаться, пока не дочитаете до конца.

Себастьян Трун, крутой парень из Google
и CEO образовательной ассоциации Udacity

Это большой-большой секрет, что статистика — штука веселая, имеет отношение к каждому из нас и способствует развитию интеллекта. Грейди Клейн и Алан Дебни разболтали его, отправив нас в путешествие по фундаментальным идеям, делающим статистику важной частью нашей жизни, в которой полным-полно самых разных фактов.

Джон Стори, профессор геномики
и статистики в Принстонском университете

ISBN 978-5-00100-260-4



9 785001 002604 >

Максимально
полезные книги на сайте
mann-ivanov-ferber.ru

издательство
МАНН, ИВАНОВ И ФЕРБЕР



facebook.com/mifbooks



vk.com/mifbooks



instagram.com/mifbooks